Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 3053 – 3073

doi:10.3906/elk-2009-27

Scale-invariant histogram of oriented gradients: novel approach for pedestrian detection in multiresolution image dataset

Sweta PANIGRAHI, Surya Narayana Raju UNDI*
National Institute of Technology Warangal, Warangal, Telangana, India

Received: 06.09.2020 • **Accepted/Published Online:** 31.07.2021 • **Final Version:** 30.11.2021

Abstract: This paper proposes a scale-invariant histogram of oriented gradients (SI-HOG) for pedestrian detection. Most of the algorithms for pedestrian detection use the HOG as the basic feature and combine other features with the HOG to form the feature set, which is usually applied with a support vector machine (SVM). Hence, the HOG feature is the most efficient and fundamental feature for pedestrian detection. However, the HOG feature produces feature vectors of different lengths for different image resolutions; thus, the feature vectors are incomparable for the SVM. The proposed method forms a scale-space pyramid wherein the histogram bin is calculated. Thus, the gradient information from all the scales is encapsulated in a single fixed-length feature vector. The proposed method is also combined with color and texture features. The proposed approach is tested on three established benchmark pedestrian datasets: INRIA, NICTA, and Daimler. An improvement of $\geq$4.5% in the miss rate is achieved for all the three datasets considered. We also show that the SI-HOG can be applied to multiresolution datasets for which the HOG feature cannot be applied. Additionally, the MapReduce model is used to obtain the same outcome. The results indicate that the proposed approach outperforms the pedestrian-detection methods considered in this work.

Key words: SI-HOG, multiresolution, pedestrian detection, interchannel voting, SVM, MapReduce

1. Introduction

In the field of computer vision, object detection in images and videos plays an imperative role. Object detection in images involves labeling the desired objects, which are specified beforehand. One recognized area of object detection is pedestrian detection. Various applications, such as human–computer interaction for video games, robotics, video surveillance, and smart vehicles have motivated research on human and pedestrian detection. Nonetheless, pedestrian detection is a challenging problem owing to the large intraclass variability arising from clothing, color, appearance, and pose. In addition, external factors such as illumination, background clutter, and partial occlusions further complicate the problem. Most pedestrian-detection algorithms involve similar stages of computation. Firstly, the pixel-level content of the image undergoes complex transformations to represent higher-level features, which are computed via feature-extraction methods. Secondly, these features for any given spatial region are fed to a classifier, which determines whether the region represents a human.

2. Literature review

Even though extensive research has been performed on pedestrian detection, significant improvements were made in recent studies, which suggests that the research has not reached a saturation point. There exists a considerable

*Correspondence: usnraju@nitw.ac.in

amount of literature related to pedestrian-detection methods. Several methods for feature extraction have been proposed, including Edge Templates [1], the Haar wavelet [2], histogram of oriented gradients (HOG) [3], the covariance descriptor [4], shape models [5], and SIFT descriptors [6].

In a major breakthrough, Dalal et al. proposed the HOG for extracting shape features. It is a dense representation of gradient information for a region. It is invariant to slight changes in translation and rotation. Local normalization helps in illumination changes. They also introduced a new annotated pedestrian dataset called INRIA with varying background and pose. Satpathy et al. [7] modified the HOG using an 18-bin histogram, yielding the extended histogram of oriented gradients (ExHOG). ExHOG solves an issue of HOG wherein gradients of opposite directions in the same cell are assigned to the same histogram bin. Both linear and nonlinear kernel support vector machines (SVM) were used and ExHOG with the nonlinear kernel performed better for the INRIA and Daimler datasets. Nigam et al. [8] proposed a multiresolution approach for detecting pedestrians using LBPs. This approach has two limitations: 1) it can be applied to only grayscale images and 2) images must be resized to a fixed size scale. Consequently, the multiresolution property of the dataset is lost. Overett et al. [9] introduced a multiresolution dataset (NICTA) of $>$25,551 pedestrians, which gives a total of 50,000 pedestrians, including left and right reflection. The negative set was sampled from 5207 high-resolution people-free images. However, the authors focused on single-resolution image sets. Yan et al. [10] proposed a multiresolution approach for traffic scenes. It employs a deformable part model [11] to map low-resolution and high-resolution pedestrians onto a common space. The detector then learns from these mapped features of different resolutions. However, correctly identifying true or false positives requires vehicle–pedestrian localization, assuming that in traffic scenes, pedestrians are around vehicles. Thus, more complexity is introduced to the system. Dollar et al. [12] evaluated the state-of-the-art pedestrian-detection methods. Detection was performed at three scales: far, medium, and near. The images were captured using a camera mounted on a vehicle. There was visible degradation for the far and near scales. Hurney et al. [13] combined HOG with a texture feature, i.e. the local binary pattern (LBP), for grayscale pedestrian images. Feature vectors of LBP variations with 16 & 8 neighborhoods and radii of 2 & 1, respectively, were obtained. The feature vectors were given to a radial basis function (rbf)-kernel SVM. The results indicated that the HOG with an LBP having neighborhood 8 and radius 1 outperformed others. Bilal et al. [14] used integer-only features from color information and orientation histograms. Classification is done by implementing a soft cascade for fast evaluation of kernel classifier. The authors could identify true negatives at the early stages from the kernel function's energies. Lahmyed et al. [15] proposed to use both visible and thermal image of a scene in pedestrian detection system. They used a modification of the OTSU method to segment the thermal images in order to get the locations of probable pedestrians. The locations get mapped to visible images, thereafter features are calculated. Bastian BT et al. [16] proposed to merge data specific dictionary learned histogram of sparse codes and aggregate channel features for pedestrian detection. Kumar et al. [17] proposed to combine histogram of significant gradients, a variation of HOG with nonredundant uniform local binary pattern to yield a feature descriptor. The authors then used a linear SVM classifier for feature training. Zhang et al. [18] proposed a pedestrian detection method that combines HOG features with the color image's edge features on depth images. They used shearlet transform to yield edge features from the color images. The combined feature is used to train the SVM classifier. Thus, the existing methods are unable to bridge the gap between the current performance and the preferred performance. Park et al. [19] proposed a deformable part-based model for detecting large-scale objects and a rigid template for small-scale objects. Context information of a scene is also utilized in the

detection process to build a context-augmented multiresolution model. However, the context cues are more effective in small-scale objects rather than large-scale ones.

To increase the performance, currently deep learning technologies are used to get a better feature representation and hence, in turn, result in better detection. Xu et al. [20] proposed a beta R-CNN which is based on 2D beta distribution to better handle occluded pedestrians. The main contribution lies in relating the full body and visible part boxes of the occluded pedestrian to assign the pixels different probability values which aids in emphasizing the visual mass center. Wu et al. [21] proposed a self-mimic learning method to detect small-scale pedestrians. A deep CNN is trained by a mimic loss which enhances the small-scale pedestrian representation by mimicking the richer feature representation of the large-scale pedestrians. Song et al. [22] proposed a progressive refinement network (PRNet) for handling occlusion in pedestrian detection. PRNet calibrates the visible-part anchors to generate full-body templates by following occlusion statistics. The authors also incorporated occlusion loss and a receptive field backfeed to generate various receptive fields. Lin et al. [23] proposed a deep feature learning system where suitable convolutional layers of varying receptive field sizes are utilized to facilitate multiscale detection. A multiscale pedestrian attention guides the model to focus on the pedestrians in the image. Karg et al. [24] described the pedestrian detection problem and the challenges it poses in driver assistance systems. Evaluation parameters and various datasets are also discussed. The authors focused on CNN based pedestrian detection.

The method proposed in this work adapts feature-extraction methods for different-scale datasets without losing the multiresolution property. This is essential in real-world applications, as same-scale images are not accessible in all scenarios. Our method achieves the following two objectives. 1) It can process multiresolution images. It extracts shape features by introducing a scale-space in HOG. Weighted gradient information of the scale-space ensures that the HOG feature vector is independent of resolution. 2) Previously, there has been no fusion of shape features, texture features, and color features with the SVM classifier. Therefore, we combine the proposed method, i.e. scale-invariant histogram of oriented gradients (SI-HOG), which gives the scale-independent shape feature, with texture and color features. As shape features are the basis for pedestrian detection, they are concatenated with texture and color features. Experiments are conducted to determine which combinations yield the best performance.

2.1. Shape features

Visual characteristics of a region or object's shape includes particulars about its boundary. Shape feature descriptors encompass edge magnitude and direction, giving it a quantitative value. HOG is widely used to extract shape feature. An introduction to HOG and its variants, HOG-18 [7] and extended-HOG (Ex-HOG) [7] are specified in the proceeding section.

Histogram of oriented gradients (HOG): HOG [3] is a dense feature-extraction method. It gives information for shape and is most popularly used for pedestrian detection [25, 26]. The process of obtaining the HOG features is described in Algorithm 1.

Algorithm 1:

1. Consider an 8 $\times$ 8 cell size from an input image.
2. Compute the vertical (g_y) and horizontal (g_x) gradients over this cell by using [–1,0,1] and $[1,0,1]^T$ filters, respectively.
3. Obtain the magnitude (M) and direction (θ) for each pixel by using equations 1 and 2.
4. Take a block size with 2 $\times$ 2 cells from the original image.

5. Obtain the 1 × 9 histogram bin (H) for each cell of the block by considering M and θ using equations 3 and 4.
6. Concatenate the four 1 × 9 histogram bins of the whole block to form a 1 × 36 histogram bin.
7. Normalize the histogram bin obtained in step 6.
8. With a stride of 1, perform steps 2 to 7 for all the blocks in the image.
9. Concatenate all these histogram bins to obtain the feature vector for the whole image.

$$M(i,j) = \sqrt{g_x^2 + g_y^2} \tag{1}$$

$$\theta(i,j) = \arctan\left(\frac{g_y}{g_x}\right) \tag{2}$$

$$H\left(\left\lceil\frac{\theta(i,j)}{20}\right\rceil \mod 9 \times 20\right) = H\left(\left\lceil\frac{\theta(i,j)}{20}\right\rceil \mod 9 \times 20\right) + (M(i,j) \times f_1) \tag{3}$$

$$H\left(\left\lceil\frac{\theta(i,j)}{20}\right\rceil \mod 9 \times 20\right) = H\left(\left\lceil\frac{\theta(i,j)}{20}\right\rceil \mod 9 \times 20\right) + (M(i,j) \times f_2) \tag{4}$$

here, $f_1 = \frac{\theta(i,j) \mod 20}{20}$ and $f_2 = 1 - f_1$.

Histogram of oriented gradients (18 bins): The difference between HOG using a 9-bin histogram (described in the previous section) and that using an 18-bin histogram is that for a cell, we obtain an 18-bin histogram ($0°$, $20°$, $40°$, ..., $340°$) as the direction is calculated in four-quadrant inverse tangent format.

Extended histogram of oriented gradients (ExHOG): Extended histogram of oriented gradients [7] is obtained from the difference and sum of HOG-18 bins.

2.2. Color features

Color features are calculated on various color models. The color model gives measurable value to colors, in the form of a tuple, having three or four values; showing the ratio in which color components are used. In the proceeding section, color feature extraction algorithms are presented on hue, saturation, intensity (HSI) color model.

(a) Color autocorrelogram Color correlogram [27] gives the joint probability of occurrence of all possible pixel levels, which results in a feature matrix of size $N \times N$ denoted as C_k. To reduce feature vector length, color autocorrelogram was proposed [27, 28] This color feature concentrate only on cooccurrence of the same color, yields a feature vector of length N, which are nothing but the diagonal values of color correlogram matrix C_k.

(b) Interchannel voting Kanaparthi et al.[1] proposed a new color feature to acquire the relationship among all the color channels, named as interchannel voting.

[1] Kanaparthi SK, Raju USN, Shanmukhi P, Aneesha GK, Rahman MEU. Image retrieval by integrating global correlation of color and intensity histograms with local texture features. Multimedia Tools and Applications 2019. doi: 10.1007/s11042-019-08029-7

2.3. Texture features

Although no formal definition exists, the texture feature is defined as a feature consisting of mutually related elements. It provides a measure of properties such as smoothness, coarseness, and regularity, which play a vital role in many image-processing applications. Different local patterns, such as LBP [29], ULBP [29], CS_LBP [30], LEP [31], DSP [32], LDP [33], LTrP [34], and ILBP [35], are used to extract texture information from an image.

2.4. Support vector machine

The SVM [36] is one of the most widely used mechanisms for solving pattern-classification problems. It is a supervised learning method that maximizes the margin of a linear decision boundary (hyperplane), thus achieving maximum separation between the two object classes. For pedestrian classification, linear and nonlinear SVM classifiers have been used in combination with various feature sets [37–39].

2.5. Big image data processing (BIDP)

Big data [40] is an encompassing term for any collection of datasets that are massive in scale, diverse, and complex. Initially, big data was defined according to the "three Vs": volume, velocity, and variety. Now, people discuss seven Vs; the other four Vs are veracity, value, visualization, and variability [41]. The authors in [42] have witnessed the increase from thousands of images to billions of images in the past 20 years, which has resulted in BIDP. In this work, we used MapReduce (MR) in MATLAB integrated with the Hadoop distributed file system (HDFS).

3. Proposed methodology

To perform pedestrian detection, first, feature extraction is performed. Because we are focusing on pedestrians, the most important feature is shape. For extracting shape information, the best-performing feature is HOG. A dataset containing images of different resolutions (dimensions) produces HOG feature vectors of different lengths, because HOG features are dependent on the size of the image. This becomes an obstacle for the classifier. If we wish to apply HOG, all the images in the dataset must be resized to the same size. Hence, the multiresolution property is lost. In the proposed method, we address this problem by utilizing a scale-space pyramid to extract shape features, which are not dependent on the size of the image. The process is described in Algorithm 2 and Figure 1 (note: The symbol '$*$' represents multiplication and the symbol '$\times$' represents that its left and right operands are width and height respectively of a matrix).

Algorithm 2: The algorithm is divided into two parts, i.e. Part I and II.

Part I (feature extraction)

1. Consider three scales to construct a pyramid structure by placing the image at the bottom ($Scale1$) and then subsampling it with a factor of 2 to create another image ($Scale2$). The process is repeated on Scale 2 to obtain the third image ($Scale3$). For Scale 1, take the largest dimension R_n, the maximal resolution from the set, having a resolution, let us say, p $\times$ q. The resolution for Scale 2 is p/2 $\times$ q/2, and that for Scale 3 is p/4 $\times$ q/4.
2. For each image, a scale-space (S_j, $1 \leq j \leq 3$) pyramid is constructed with the three resolutions, as described in Step 1.
3. For Scale 1, consider the cell size as m $\times$ m. For Scale 2, the cell size is m/2 $\times$ m/2 and for Scale 3, it

is m/4 × m/4. Thus, the same number of cells is maintained for all the scales.

4. Obtain a 9-bin histogram (H) using equations 3 and 4 for every cell in a scale S_j, with $1 \leq j \leq 3$.
5. Concatenate the histogram bins obtained in Step 4 for each scale S_j to form $H_{scale,j}$, where $1 \leq j \leq 3$.
6. Take the average of $H_{scale,1}$, $H_{scale,2}$, and $H_{scale,3}$ according to equation 5.
7. Considering 2 × 2 overlapping cells with a stride of 1, apply block normalization on H_{avg} to obtain the final feature vector for the image.

$$H_{avg} = \frac{\sum_{j=1}^{3} H_{scale,j}}{3} \tag{5}$$

where the size of $H_{scale,j} = 9 \times \frac{p*q}{m*m}$

Part II (Classification)

1. The feature vectors of both positive and negative images (with labels 1 and –1, respectively) are used to train the SVM.
2. The test set feature vector including the positive and negative images is given to the trained SVM model to obtain a label (either 1 or –1).
3. The actual and predicted labels of the test set are compared to obtain a confusion matrix.

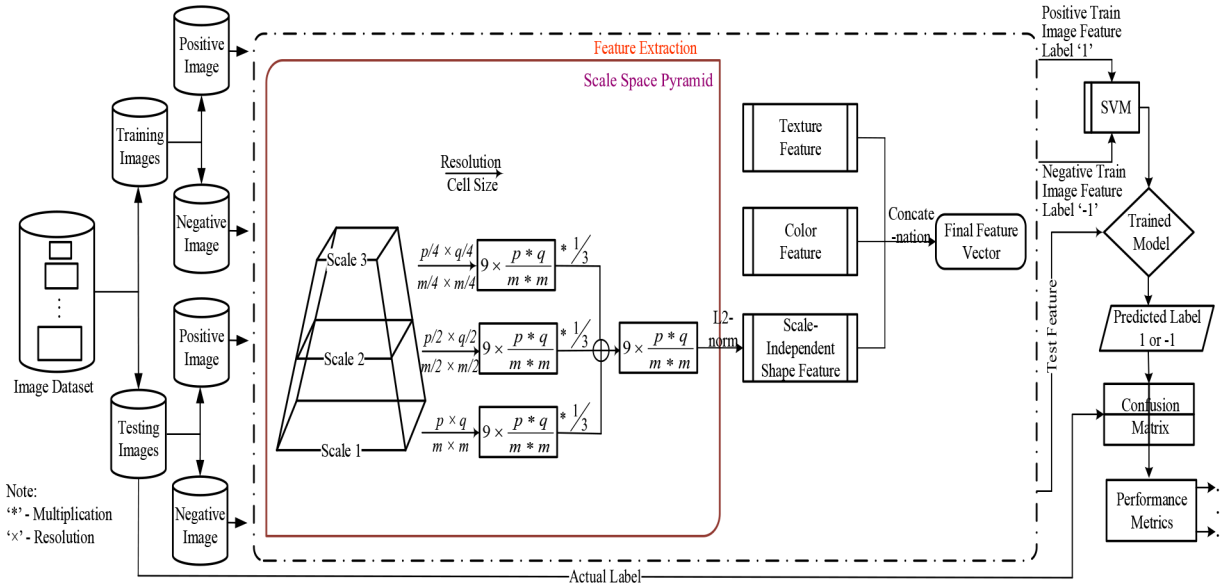


Figure 1. Proposed SI-HOG method. Scale-invariant feature extraction of training and testing images for calculation of performance metrics.

In Figure 1, the image dataset is divided into two parts, i.e. training and testing images. Both the Training and testing images are subdivided into positive and negative images. On these images, Algorithm 2's Part 1 (feature extraction) is applied to obtain SI-HOG. The size of the SI-HOG feature vector obtained is 9 × $\frac{p*q}{m*m}$ from the scale space pyramid. The mathematical proof is given below:

1. The concatenated Histogram ($H_{scale,j}$) of all the scales (S_j) is formed as:

   1.1. Scale 1: The resolution of image is p × q and the cell size is m × m. Therefore, we get $\frac{p*q}{m*m}$

number of cells. For each cell, a 9-bin histogram is formed. Thus, for scale 1, 9 $\times$ $\frac{p*q}{m*m}$ $(=H_{scale,1})$ is yielded by concatenating each 9-bin histogram vertically.

1.2. Scale 2: The resolution of image is p/2 $\times$ q/2 and the cell size is m/2 $\times$ m/2. Therefore, we get a $\frac{p/2*q/2}{m/2*m/2}$ number of cells which results in $\frac{p*q}{m*m}$ cells. Similarly, for each cell, a 9-bin histogram is formed. Finally, for scale 2 also, 9 $\times$ $\frac{p*q}{m*m}$ $(=H_{scale,2})$ is yielded by concatenating each 9-bin histogram vertically.

1.3. Scale 3: The resolution of image is p/4 $\times$ q/4 and the cell size is m/4 $\times$ m/4. Therefore, we get a $\frac{p/4*q/4}{m/4*m/4}$ number of cells which results in $\frac{p*q}{m*m}$ cells. Similarly, for each cell, a 9-bin histogram is formed and 9 $\times$ $\frac{p*q}{m*m}$ $(=H_{scale,3})$ is yielded by concatenating each 9-bin histogram vertically.

2. As all the histogram of the scales are given equal importance, $H_{scale,1}$, $H_{scale,2}$ and $H_{scale,3}$ each are multiplied by a factor of $\frac{1}{3}$ and then summed to get H_{avg}. The size of H_{avg} will be equal to that of $H_{scale,1}$, $H_{scale,2}$ and $H_{scale,3}$, i.e. 9 $\times$ $\frac{p*q}{m*m}$.
3. As per the construction of a block in Figure 1 comprising of four cells, H_{avg} is block normalized following L2-norm to yield the final feature vector. The length of the final feature vector is given in equation 6.

$$Length\ of\ SI\text{-}HOG\ feature\ vector = number\ of\ blocks * length\ of\ each\ block \tag{6}$$

where number of blocks = ($\frac{p}{m}$-1) $*$ ($\frac{q}{m}-1$) and length of each block = number of cells per block $*$ number of bins = $4 * 9 = 36$.

The feature vector yielded from positive training set is assigned label '1' and that of negative training set is assigned label '-1'. The feature vector along with the label vector from the training set is used to train SVM, which yields a model. The positive testing set and the negative testing set along with their actual labels are given as input to the SVM model. The output is a set of predicted labels which is further used to construct confusion matrix and the performance metrics

We used the state-of-the-art MapReduce (MR) paradigm to obtain the final results. In general, this paradigm can be used for processing a large number of images. In our proposed method, we used two MR Jobs. MR Job1 was used to convert the image dataset into sequence files. Sequence file is a native MapReduce data structure which is formed by the concatenation of (key, value) pairs. As image datasets have small images, sequence files group them to form a large file which is suitable for MapReduce. It is primarily used in MapReduce as input/output formats. Sequence files for Job1 were formed by concatenating individual images as key and the pixel values of images as value. The MR Job2 took the sequence files and calculated the feature vector of all the images. For Job2, the output was a (key, value) pair comprising only the image number and feature vector. The feature vectors were given to the SVM in the training phase, which provides the model. The same process was applied for the testing image dataset and then the confusion matrix was obtained, whereby we evaluated the performance of the proposed method. The entire process is shown in Figure 2.

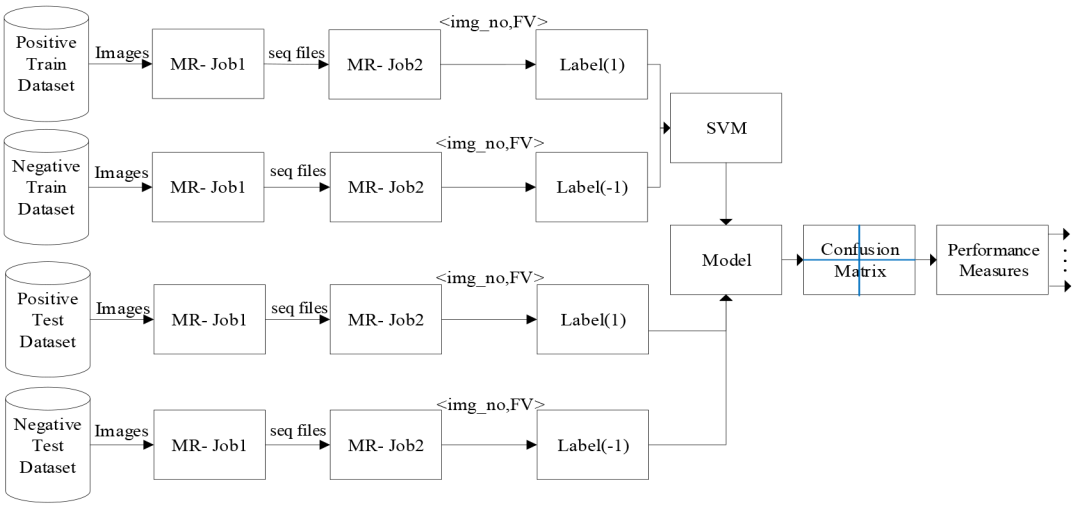


Figure 2. MR paradigm for the proposed SI-HOG. Computation of scale-invariant feature extraction of training and testing for calculation of performance metrics by using MapReduce paradigm.

4. Results and discussions

We compared the proposed method with existing algorithms for three standard pedestrian datasets: INRIA[1], NICTA[2], and Daimler[3]. For the three datasets, per-window evaluation was employed. The value of m, which determines the cell size for the scale-space, was 16, 8, and 12 for INRIA, NICTA, and Daimler, respectively. These values were experimentally determined to perform well for the datasets. For INRIA, a linear SVM classifier with a regularization parameter (C) of 0.01 was used, whereas for NICTA and Daimler, a nonlinear SVM classifier with an rbf kernel was used. The classification was performed using the LIBSVM machine-learning library. From the SVM, the confusion matrix was obtained, as explained in Subsection 4.1. Then, performance metrics were obtained, as given in Subsection 4.2. Details regarding the dataset usage and properties are presented in Subsection 4.3. Subsection 4.4 discusses the performance for the INRIA, NICTA, and Daimler pedestrian datasets. Subsection 4.5 describes the Friedman test analysis of the results. Subsection 4.6 describes the implementation using the MR model.

4.1. Confusion matrix

For binary classification, i.e. two-class (person or nonperson) classification, the SVM generates a predicted label of 1 (positive/person) or –1 (negative/nonperson). Given the actual label of the test set, a 2 × 2 confusion matrix is formed. From the confusion matrix, the following resulting particulars are obtained.

True Positive(TP): A pedestrian is present in the test image and the image is classified as positive.

False Positive(FP): No pedestrian is present in the test image and the image is classified as positive.

False Negative(FN): A pedestrian is present in the test image and the image is classified as negative.

True Negative(TN): No pedestrian is present in the test image and the image is classified as negative.

[1]INRIA Pedestrian Dataset (2005). Dalal N, Triggs B. [online]. Website: http://pascal.inrialpes.fr/data/human/ [accessed 07 August 2020].

[2]NICTA Pedestrians Dataset (2008). Overett G, Petersson L, Brewer N, Andersson L, Pettersson N. [online]. Website: https://data.csiro.au/collections/#collection/CIcsiro:23454v1 [accessed 07 August 2020].

[3]Daimler Mono Ped. Classification Benchmark Data Set (2006). Munder S, Gavrila DM. [online] Website: http://www.lookingatpeople.com/download-daimler-ped-class-benchmark/index.html [accessed 07 August 2020].

4.2. Detection metrics [43, 44]

On the basis of the four values from the confusion matrix, the receiver operator characteristics (ROC) and the detection error tradeoff(DET) curve are obtained.

Receiver operator characteristics (ROC) curve [45]: The ROC curve is a two-dimensional graph formed by plotting the false-positive rate (FPR) on the X-axis against the true-positive rate (TPR) on the Y-axis.

Detection error tradeoff (DET) curve [46]: The DET curve is a log–log curve with the error rates on both axes, giving uniform treatment to both types of error, i.e. false positive and false negative.

4.3. Dataset properties

To evaluate the performance of the proposed method, we used three standard pedestrian datasets: INRIA, NICTA, and Daimler. The INRIA pedestrian dataset contained images of pedestrians against a wide variety of backgrounds, including crowds. The images were available in normalized windows of size 64 × 128. The same size was maintained while extracting negative windows. Bicubic interpolation was used to form the multiresolution datasets. The NICTA pedestrian dataset contained both positive and negative pedestrian images having six resolutions: 8 × 20, 16 × 20, 16 × 40, 32 × 40, 32 × 80, and 64 × 80. Thus, we did not need to create multiresolution images, as they were already provided. The Daimler pedestrian dataset contained images with a size of 18 × 36 extracted from videos captured in various settings to avoid biases in clothes and pose. Similar to the case of INRIA, we obtained multiresolution images via resizing. Both NICTA and Daimler provided additional person-free images, i.e. negative images, for bootstrapping. The characteristics of the INRIA, NICTA, and Daimler datasets are presented in Table 1.

Table 1. Characteristics of the INRIA, NICTA, and Daimler pedestrian datasets.

Dataset	Resolution	Image dimension	Training		Testing	
			No. of Positive Images	No. of negative images	No. of positive images	No. of negative images
INRIA (color)	Single	64 × 128	2416	1218 * 10	1132	4530 * 10
	Multi	128 × 256	805	406 * 10	377	151 * 10
		64 × 128	805	406 * 10	377	151 * 10
		32 × 64	805	406 * 10	378	151 * 10
NICTA (color)	Single	64 × 80	12000	18000	6000	9000
	Multi	64 × 80	2000	3000	1000	1500
		32 × 80	2000	3000	1000	1500
		32 × 40	2000	3000	1000	1500
		16 × 40	2000	3000	1000	1500
		16 × 20	2000	3000	1000	1500
		8 × 16	2000	3000	1000	1500
Daimler (gray)	Single	18 × 36	4800	24000	2400	12000
	Multi	72 × 144	1600	8000	800	4000
		36 × 72	1600	8000	800	4000
		18 × 36	1600	8000	800	4000

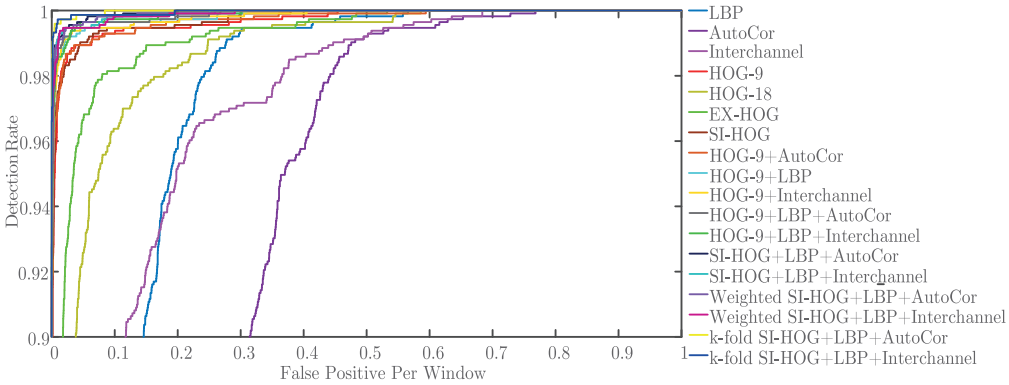
4.4. Performance metrics for different image datasets

The ROC curves for the INRIA single-resolution and multiresolution datasets are presented in Figures 3a and 3b, respectively. The proposed method exhibited the best performance among the feature-extraction methods tested. Here, we have shown ROC curves for only one dataset. The DET curves for the INRIA, NICTA, and Daimler single-resolution and multiresolution datasets are shown in Figures 4a and 4b, Figures 5a and 5b, and Figures 6a and 6b, respectively. For the grayscale-image dataset Daimler, the miss rate values at 10^{-2} are listed in Table 2, on the basis of the DET curves shown in Figure 6. And according to the DET curves shown in Figures 4 and 5, the miss rate values at 10^{-3} are presented in Table 3 for the single-resolution and multiresolution images from both the color-image datasets: INRIA and NICTA. In the single-resolution case, the proposed method achieved the lowest miss rate for all three datasets, i.e. INRIA, NICTA, and Daimler (7.47%, 20.56%, and 33.20%, respectively). The second lowest miss rate was achieved by HOG-9. Thus, the proposed method was the best-performing method when compared with other shape features as well as texture and color features. To depict an object better, texture (LBP), and color (autocorrelogram & interchannel) features are combined with HOG-9 and SI-HOG. For INRIA and NICTA, combining texture and color features with the proposed method further improved the miss rate. In contrast, because Daimler was a grayscale dataset, only texture features were employed in the proposed method. In the case of Daimler, significant changes in the miss rate were not observed. In the multiresolution case, as HOG-9 and other variants of HOG could not be applied, the proposed method was the only shape feature-extraction method. With the addition of texture and color information, miss rates of 6.98% and 40.24% were achieved for INRIA and NICTA, respectively. For Daimler, the addition of texture information to SI-HOG yielded a miss rate of 33.21%. Therefore, SI-HOG was the best-performing method with the lowest miss rate and the performance was improved via the addition of color and texture information. Table 4 shows comparison of the miss rates of the proposed method with the existing multiresolution methods.

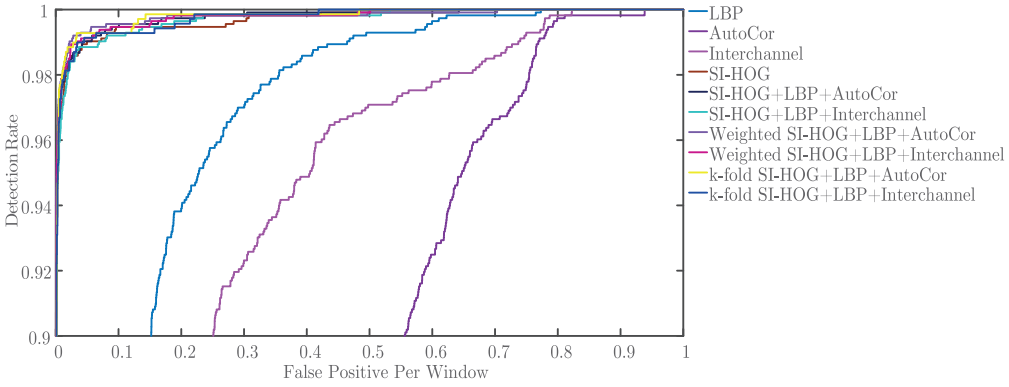
Results of fusion strategies: Two fusion strategies namely weighted and k-fold SI-HOG+LBP+AutoCor and SI-HOG+LBP+Interchannel are evaluated for INRIA, NICTA, and Daimler single- and multiresolution datasets. The miss rates for Daimler datasets are shown in Table 2 and for INRIA and NICTA datasets are shown in Table 3. In the case of weighted methods, the weight was taken with respect to the ratio of (TP+TN) to (TP+FN+FP+TN) of each method. The weight was multiplied by their respective train and test feature vectors to form weighted feature vectors. In the case of k-fold, the total of the train-test set was divided into k parts where k = 5. Then the performance was evaluated on the train:test set in the ratio of 4:1 parts. In the case of weighted fusion, there is an improvement in miss rate in the single and multiresolution INRIA dataset and in the single resolution NICTA dataset. In the case of k-fold fusion, there is a significant improvement of miss rate in the single resolution INRIA and the multiresolution NICTA datasets and also in both the single and multiresolution Daimler datasets. The fusion methods are reflected in their respective DET curves.

4.5. Friedman test analysis

Friedman test [47] is significantly used in the statistical analysis used in research studies. This test with a significance level of 95% and $\alpha = 0.05$ was conducted to the miss rates yielded by each of the algorithms. In this work, there are 14 and 6 independent variables in the case of single- and multiresolution datasets, respectively. Including the fusion strategies, there are 18 and 10 independent variables in the case of single- and multiresolution datasets, respectively. The rank of each of these variables are examined using the Friedman test. The null hypothesis states that all the algorithms are equivalent. The mean rank produced by Friedman test



(a)

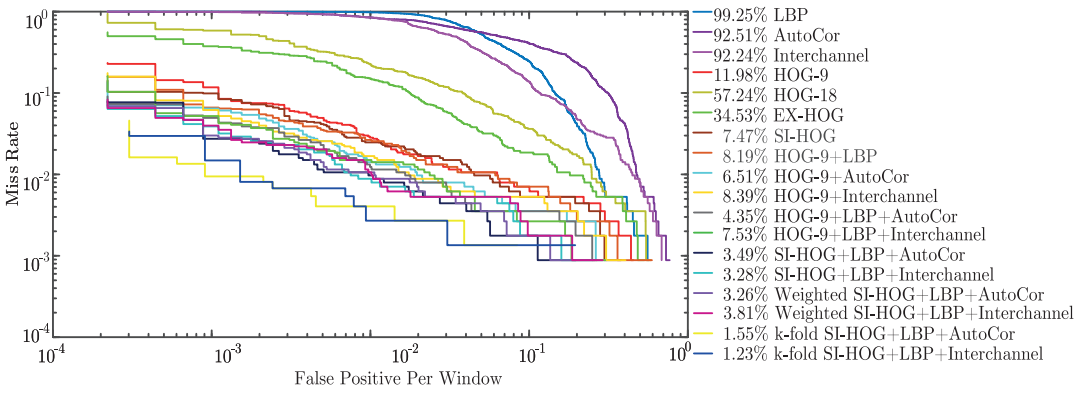


(b)

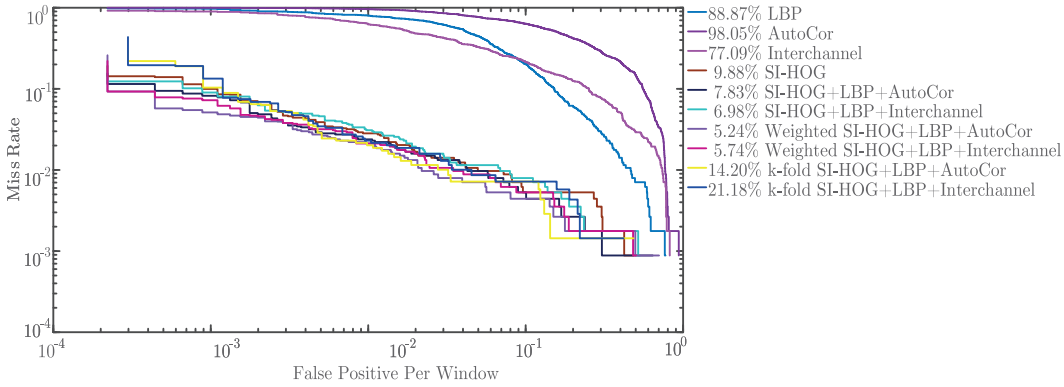
Figure 3. Receiver operator characteristic curve for INRIA pedestrian dataset. This graph shows the false positive per window vs detection rate curve for (a) single resolution images of the dataset. (b) Multiresolution images of the dataset.

Table 2. Miss rates for single-resolution and multiresolution images from the Daimler pedestrian dataset at 0.01 false positives per window.

Method	Miss rate (%)	
	Daimler	
	Single-resolution	Multiresolution
LBP	48.74	69.38
HOG-9	37.89	NA
HOG-18	53.12	NA
Ex-HOG	49.39	NA
SI-HOG(Proposed)	33.20	32.93
HOG-9+LBP	36.60	NA
SI-HOG+LBP	32.67	33.31
Weighted SI-HOG+LBP	33.17	34.06
k-fold SI-HOG+LBP	13.98	12.87



(a)



(b)

Figure 4. Detection error tradeoff curve for INRIA pedestrian dataset. This graph shows the false positive per window vs miss rate in log-log scale for (a) Single-resolution images of the dataset. (b) Multiresolution images of the dataset.

for each of the pedestrian algorithm is given in Table 5. The proposed method SI-HOG has the highest rank when compared with single methods, i.e. among LBP, AutoCor, Interchannel, HOG-9, HOG-18, Ex-HOG, and SI-HOG. The proposed concatenation, i.e. SI-HOG+LBP+AutoCor and SI-HOG+LBP+Interchannel has the first and second ranks, respectively, in the case of single resolution. In the case of multiresolution, both have the first rank. In the case of single resolution, the calculated chi-square is 25.557. The critical value of chi-square at a degree of freedom (k-1) 13 is 21.026; where k is the number of algorithms. As the calculated value of chi-square is greater than the critical value of chi-square, the null hypothesis is rejected. In the case of multiresolution, the calculated chi-square is 9.428. The critical value of chi-square at degree of freedom 5 is 11.070. As the calculated chi-square is less than the critical chi-square, the null hypothesis is failed to be rejected. Iman et al. [48] showed that Friedman's chi-square is undesirably conservative and gave a better statistic which is distributed according to the F-distribution with k-1 and (k-1)(N-1) degrees of freedom; where k and N are the number of algorithm and datasets, respectively. On application of F-distribution on the multiresolution datasets, the calculated F-distribution is 16.482. The critical value of F-distribution with degree of freedom (5,5) and $\alpha = 0.05$ is 5.05. As the calculated F-distribution is greater than the critical F-distribution, the null hypothesis is rejected.

*Friedman test analysis including the fusion strategies:*The Friedman test is conducted on the two fusion strategies

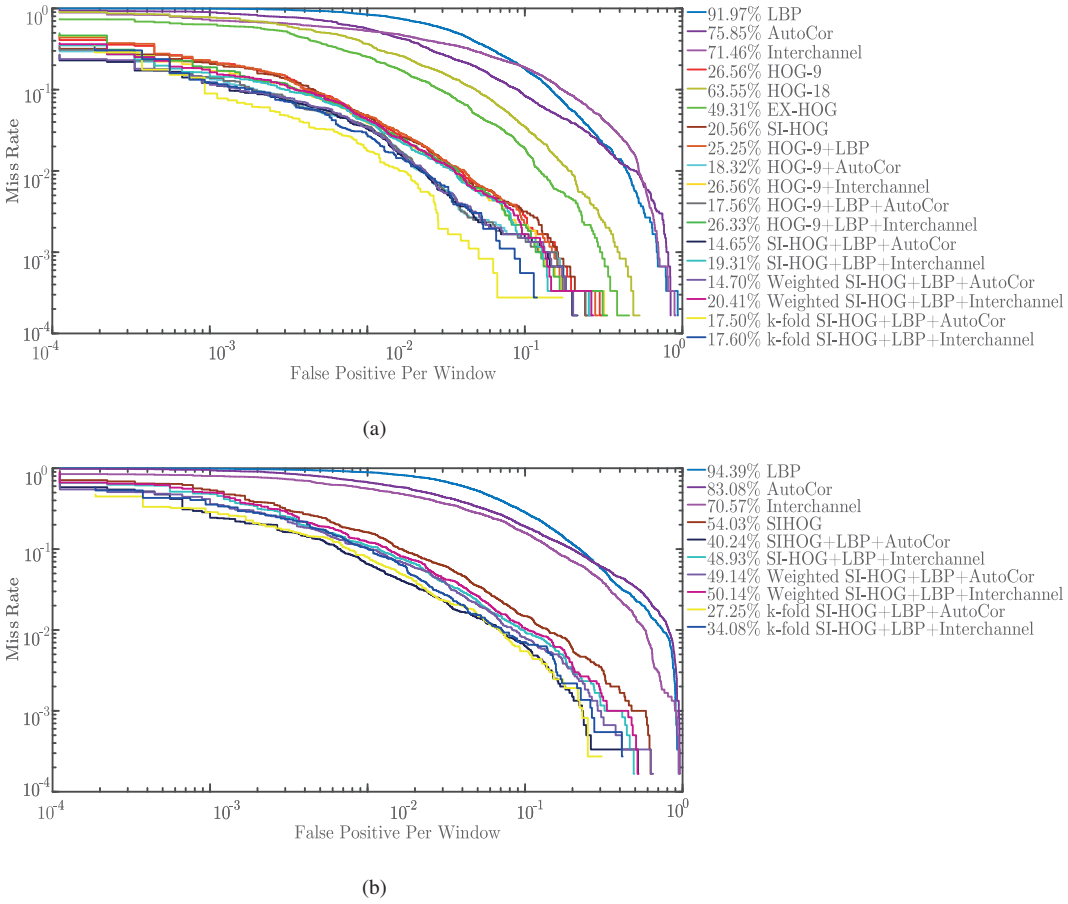


Figure 5. Detection error tradeoff curve for NICTA pedestrian dataset. This graph shows the false positive per window vs miss rate in log-log scale for (a) Single resolution images of the dataset. (b) Multiresolution images of the dataset.

namely weighted and k-fold SI-HOG+LBP+AutoCor and SI-HOG+LBP+Interchannel. In the case of single resolution INRIA and NICTA datasets, there are 18 algorithms whose Mean Rank yielded is [18, 17, 16, 12.75, 15, 14, 9, 10.5, 7, 12.25, 5.5, 10.5, 3, 5.5, 2.5, 7, 2.5, 3] as per the sequence in Table 3. It can be observed that except for the weighted SI-HOG+LBP+Interchannel all the 5 proposed methods are with high mean rank. The calculated chi-square is 32.88. The critical value of chi-square at a degree of freedom (k-1) 17 is 27.59; where k is the number of algorithms. As the calculated value of chi-square is greater than the critical value of chi-square, the null hypothesis is rejected. In the case of multiresolution INRIA and NICTA datasets, there are 10 algorithms whose mean rank yielded [9.5, 9.5, 8, 6, 3.5, 3.5, 3, 4, 3.5, 4.5] as per the sequence in Table 3. In this case, the 6 proposed methods are with top ranks as well. The calculated chi-square is 13.09. The critical value of chi-square at a degree of freedom 9 is 16.919. As the calculated chi-square is less than the critical chi-square, the null hypothesis is failed to be rejected. On application of F-distribution on the multiresolution datasets, the calculated F-distribution is 2.665. The critical value of F-distribution with degree of freedom (9,9) and $\alpha = 0.05$ is 3.17. There is rejection of null hypothesis here as well. It may be concluded that this is due to, among the 10 methods considered, the 6 fusions of SI-HOG yielded similar mean ranks.

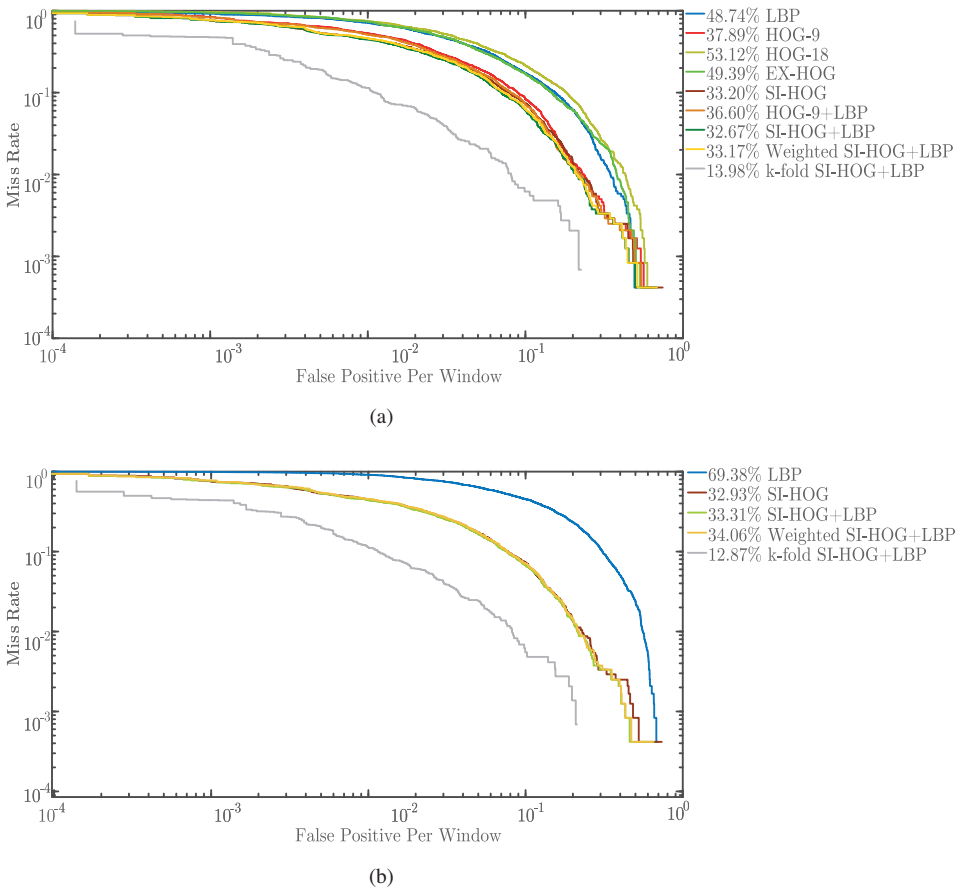


Figure 6. Detection error tradeoff curve for Daimler pedestrian dataset. This graph shows the false positive per window vs miss rate in log-log scale (a) Single resolution images of the dataset. (b) Multiresolution images of the dataset.

4.6. Complementarity and diversity analysis

The measures complementarity and diversity are used to quantify the success in ensembling features in this work. Through complementarity analysis, the superiority of the fusion strategies can be determined. Complementarity analysis can be done by counting the number of TPs. In this work, the TN is mentioned as well. This gives us a better understanding of the hit rates of the ensemble feature strategies. For the INRIA, NICTA, and Daimler single and multiresolution datasets, these values are reflected in Tables 6–8, respectively. It can be observed that the proposed SI-HOG+LBP+AutoCor and SI-HOG+LBP+Interchannel along with its fusion strategies i.e. weighted and k-fold method amount to a greater quantity of TPs and TNs. Diversity analysis is an important factor in ensemble methods. It shows that even when the miss rate of two or more methods can be similar, the overlap of TPs can be different, which serves as an essential measure for fusion strategies. Nonpairwise diversity is quantified by the interrater agreement measure κ [49] given in equation 7. The pairwise diversity is calculated by Q statistics[50] given in equation 8. The pairwise and nonpairwise diversity of INRIA, NICTA and Daimler single and multiresolution datasets are given in Tables 6–8, respectively. It can be observed that the fused models and the fusion strategies involving the proposed method are giving varying levels of diversity. The fusion of the methods increases the TPs and decreases the false negatives (FNs) in the ensemble model. Below an example for both κ and Q statistics calculation is given.

Table 3. Miss rates for single-resolution and multiresolution images from the INRIA and NICTA pedestrian datasets at 0.001 false positives per window.

Method	Miss rate (%)			
	INRIA		NICTA	
	Single-resolution	Multi-resolution	Single-resolution	Multi-resolution
LBP	99.25	88.87	91.97	94.39
AutoCor	92.51	98.05	75.85	83.08
Interchannel	92.24	77.09	71.46	70.57
HOG-9	11.98	NA	26.56	NA
HOG-18	57.24	NA	63.55	NA
Ex-HOG	34.53	NA	49.31	NA
SI-HOG(Proposed)	7.47	9.88	20.56	54.03
HOG-9+LBP	8.19	NA	25.25	NA
HOG-9+AutoCor	6.51	NA	18.32	NA
HOG-9+Interchannel	8.39	NA	26.56	NA
HOG-9+LBP+AutoCor	4.35	NA	17.56	NA
HOG-9+LBP+Interchannel	7.53	NA	26.33	NA
SI-HOG+LBP+AutoCor	3.49	7.83	14.65	40.24
SI-HOG+LBP+Interchannel	3.28	6.98	19.31	48.93
Weighted SI-HOG+LBP+AutoCor	3.26	5.24	14.70	49.14
Weighted SI-HOG+LBP+Interchannel	3.81	5.74	20.41	50.14
k-fold SI-HOG+LBP+AutoCor	1.55	14.20	17.50	27.25
k-fold SI-HOG+LBP+Interchannel	1.23	21.18	17.60	34.08

Table 4. Comparison of miss rates for INRIA pedestrian dataset.

Methods	Miss rate (%)
Lin et al. [23]	4.7
Nigam et al. [8]	7
SI-HOG+LBP+Interchannel (Ours)	3.28

Diversity computation example of k-fold SIHOG+LBP+Interchannel for NICTA single resolution dataset:-
$SI-HOG_{TP} = 3532$; $SI-HOG_{FN} = 81$; $SI-HOG_{FP} = 96$; $SI-HOG_{TN} = 5291$
$LBP_{TP} = 3242$; $LBP_{FN} = 371$; $LBP_{FP} = 680$; $LBP_{TN} = 4707$
$Interchannel_{TP} = 3177$; $Interchannel_{FN} = 436$; $Interchannel_{FP} = 676$; $LBP_{TN} = 4711$

1. κ: The value of L = 3, N = 9000. Applying the equation 7, the value of $\bar{\rho} = 0.913$ and $\kappa = 0.529$.
2. Q_{av}: The average of $Q_{SI-HOG,LBP}$, $Q_{SI-HOG,Interchannel}$ and $Q_{LBP,Interchannel}$ is computed here. For $Q_{SI-HOG,LBP}$, $N^{11} = 7949$, $N^{00} = 177$, $N^{01} = 0$ and $N^{10} = 874$. Applying the equation in 8, $Q_{SI-HOG,LBP} = 1$. Similarly, $Q_{SI-HOG,Interchannel} = 1$ and $Q_{LBP,Interchannel} = 0.999$.

Table 5. Average ranking of the pedestrian detectors on the single and multiresolution dataset using Friedman statistical test.

Methods	Single-resolution		Multiresolution	
	Mean rank	Rank	Mean rank	Rank
LBP	14	12	5.5	4
AutoCor	13	11	5.5	4
Interchannel	12	10	4	3
HOG-9	8.75	7	NA	NA
HOG-18	11	9	NA	NA
Ex-HOG	10	8	NA	NA
SI-HOG(Proposed)	5	4	3	2
HOG-9+LBP	6.5	5	NA	NA
HOG-9+AutoCor	3.5	3	NA	NA
HOG-9+Interchannel	8.25	6	NA	NA
HOG-9+LBP+AutoCor	2.5	2	NA	NA
HOG-9+LBP+Interchannel	6.5	5	NA	NA
SI-HOG+LBP+AutoCor	1.5	1	1.5	1
SI-HOG+LBP+Interchannel	2.5	2	1.5	1

Therefore, $Q_{av} = 0.999$.

$$\kappa = 1 - \frac{\frac{1}{L}\sum_{j=1}^{N} l(z_j)(L - l(z_j))}{N(L-1)\bar{\rho}(1-\bar{\rho})} \;\; where, \;\; l(z_j) = \sum_{i=1}^{L} y_{j,i} \; and \; \bar{\rho} = \frac{1}{NL}\sum_{j=1}^{N} l(z_j) \tag{7}$$

where L is the number of detectors and N is the count of the groundtruth. The $y_{j,i}$ term refers to whether the i^{th} detector detects j^{th} ground truth or not and then assigned to 1 or 0 respectively.

$$Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k} \;\; where, \;\; Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{8}$$

where L is the number of detectors. For two detectors, i and k, the Q statistics is given by $Q_{i,k}$. The N^{11} and N^{00} are the number of hits and misses made by both the detectors respectively. N^{01} is the number of groundtruth missed by first detector and hit by second detector. The converse count is assigned to N^{10}. The total groundtruth count $N = N^{11} + N^{00} + N^{01} + N^{10}$.

4.7. Implementation using MapReduce (MR) model

We implemented the proposed method by using the MR programming model. Four different cluster setups were used, as well as MR programming with MATLAB and Hadoop. In these implementations, all the images were converted into .seq files and uploaded to the HDFS. The images were read from the HDFS and executed on a single system with the following specifications: 4 workers/12 workers parallel pool/3 node Hadoop cluster.

Table 6. Complementarity and diversity (nonpairwise and pairwise) analysis of INRIA single- and multiresolution datasets.

Fusion methods	INRIA					
	Single Resolution			Multiresolution		
	Complement -arity [TP,TN]	Diversity		Complement -arity [TP,TN]	Diversity	
		Non -pairwise κ	Pair -wise Q_{av}		Non -pairwise κ	Pair -wise Q_{av}
HOG-9+LBP	[1084,4514]	0.096	1.000	NA	NA	NA
HOG-9+AutoCor	[1091,4514]	0.004	0.899	NA	NA	NA
HOG-9+Interchannel	[1082,4512]	0.018	0.934	NA	NA	NA
HOG-9+LBP+AutoCor	[1099,4516]	–0.101	0.940	NA	NA	NA
HOG-9+LBP+Interchannel	[1099,4519]	–0.077	0.955	NA	NA	NA
SI-HOG+LBP+AutoCor	[1112,4517]	–0.106	0.950	[1093,4514]	0.452	0.920
SI-HOG+LBP+Interchannel	[1104,4520]	–0.077	0.965	[1081,4506]	0.452	0.920
Weighted SI-HOG+LBP+AutoCor	[1106,4517]	0.167	0.940	[1085,4520]	0.453	0.926
Weighted SI-HOG+LBP+Interchannel	[1104,4522]	0.167	0.940	[1084,4520]	0.453	0.926
k-fold SI-HOG+LBP+AutoCor	[733,3307]	–0.147	0.956	[671,3341]	0.457	0.889
k-fold SI-HOG+LBP+Interchannel	[730,3307]	–0.068	0.983	[667,3343]	0.457	0.889

Table 7. Complementarity and Diversity(Non-pairwise and Pairwise) Analysis of NICTA Single and multiresolution datasets

Fusion methods	NICTA					
	Single-resolution			Multiresolution		
	Complement -arity [TP,TN]	Diversity		Complement -arity [TP,TN]	Diversity	
		Non -pairwise κ	Pair -wise Q_{av}		Non -pairwise κ	Pair -wise Q_{av}
HOG-9+LBP	[5871,8803]	0.238	1.000	NA	NA	NA
HOG-9+AutoCor	[5870,8810]	0.337	1.000	NA	NA	NA
HOG-9+Interchannel	[5882,8823]	0.214	1.000	NA	NA	NA
HOG-9+LBP+AutoCor	[5912,8833]	0.377	1.000	NA	NA	NA
HOG-9+LBP+Interchannel	[5878,8817]	0.517	1.000	NA	NA	NA
SI-HOG+LBP+AutoCor	[5913,8845]	0.375	1.000	[5825,8780]	0.446	0.994
SI-HOG+LBP+Interchannel	[5891,8827]	0.515	1.000	[5747,8751]	0.480	1.000
Weighted SI-HOG+LBP+AutoCor	[5914,8841]	0.318	1.000	[5750,8738]	0.455	0.999
Weighted SI-HOG+LBP+Interchannel	[5884,8814]	0.310	1.000	[5740,8727]	0.453	1.000
k-fold SI-HOG+LBP+AutoCor	[3572,5311]	0.418	1.000	[3532,5222]	0.466	0.999
k-fold SI-HOG+LBP+Interchannel	[3561,5304]	0.529	0.999	[3525,5212]	0.413	1.000

Table 8. Complementarity and diversity (nonpairwise and pairwise) analysis of Daimler single- and multiresolution datasets.

Fusion methods	Daimler					
	Single-resolution			Multiresolution		
	Complementarity [TP,TN]	Diversity		Complementarity [TP,TN]	Diversity	
		Non -pairwise κ	Pair -wise Q_{av}		Non -pairwise κ	Pair -wise Q_{av}
HOG-9+LBP	[1313,11826]	0.528	0.960	NA	NA	NA
SIHOG+LBP	[1450,11834]	0.471	0.950	[1450,11823]	0.471	0.952
Weighted SIHOG+LBP	[1400,11838]	0.488	0.957	[1412,11834]	0.482	0.955
k-fold SIHOG+LBP	[1285,7118]	0.135	0.879	[1278,7119]	0.132	0.875

5. Conclusion

We proposed a scale-space pyramid-based shape feature-extraction method: SI-HOG. SI-HOG overcomes the shortcoming of HOG, i.e. it is not applicable to multiresolution images, by considering gradient information from different scales of an image, making it resolution-independent. Furthermore, we proposed the addition of texture and color information to SI-HOG for extracting a more detailed form of features. We evaluated the performance using three datasets, i.e. INRIA, NICTA, and Daimler, considering both single-resolution and multiresolution images. In NICTA, both single-resolution and multiresolution images were available, whereas for INRIA and Daimler, multiresolution images were created using bicubic interpolation. SI-HOG outperformed the existing LBP, AutoCor, Interchannel, HOG-9 bins, HOG-18 bins, and ExHOG methods in both the single-resolution and multiresolution cases for all three datasets. When texture and color features were added, for INRIA and NICTA, SI-HOG+LBP+Interchannel and SI-HOG+LBP+AutoCor exhibited the best performance in both the single-resolution and multiresolution cases. For Daimler, the results did not vary significantly with the addition of texture features (SI-HOG+LBP). Hence, SI-HOG is the best-performing shape feature among all the individual feature-extraction methods tested in this work. The proposed method was also tested with two fusion methods: weighted and k-fold and the results are compared. Friedman test analysis is performed on all the methods and the ensemble methods are analyzed by complementarity and diversity (nonpairwise and pairwise) computation.

References

[1] Gavrila DM. A bayesian, exemplar-based approach to hierarchical shape matching. IEEE Transactions on Pattern Analysis & Machine Intelligence 2007; 29 (8): 1408-1421. doi: 10.1109/TPAMI.2007.1062

[2] Papageorgiou CP. A trainable system for object detection in images and video sequences. PhD, Massachusetts Institute of Technology, Cambridge, MA, USA, 1991.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: International Conference on Computer Vision & Pattern Recognition; San Diego, USA; 2005. pp. 886–893.

[4] Tuzel O, Porikli F, Meer P. Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence 2008; 30 (10): 1713-1727. doi: 10.1109/TPAMI.2008.75

[5] Leibe B, Leonardis A, Schiele B. Robust object detection with interleaved categorization and segmentation. International Journal of Computer Vision 2008; 77 (1-3): 259-289. doi: 10.1007/s11263-007-0095-3

[6] Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 2004; 60 (2): 91-110. doi: 10.1023/B:VISI.0000029664.99615.94

[7] Satpathy A, Jiang X, Eng HL. Human detection by quadratic classification on subspace of extended histogram of gradients. IEEE Transactions on Image Processing 2013; 23 (1): 287-297. doi: 10.1109/TIP.2013.2264677

[8] Nigam S, Khare A. Multiresolution approach for multiple human detection using moments and local binary patterns. Multimedia Tools and Applications 2015; 74 (17): 7037-7062. doi: 10.1007/s11042-014-1951-0

[9] Overett G, Petersson L, Brewer N, Andersson L, Pettersson N. A new pedestrian dataset for supervised learning. In: IEEE Intelligent Vehicles Symposium; Eindhoven, Netherlands; 2008. pp. 373-378.

[10] Yan J, Zhang X, Lei Z, Liao S, Li SZ. Robust multi-resolution pedestrian detection in traffic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition; Portland, USA; 2013. pp. 3033-3040.

[11] Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 2009; 32 (9): 1627-1645. doi: 10.1109/TPAMI.2009.167

[12] Dollar P, Wojek C, Schiele B, Perona P. Pedestrian Detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 2012; 34 (4): 743-761. doi: 10.1109/TPAMI.2011.155

[13] Hurney P, Waldron P, Morgan F, Jones E, Glavin M. Night-time pedestrian classification with histograms of oriented gradients-local binary patterns vectors. IET Intelligent Transport Systems 2014; 9 (1): 75-85. doi: 10.1049/iet-its.2013.0163

[14] Bilal M, Hanif MS. High performance real-time pedestrian detection using light weight features and fast cascaded kernel SVM classification. Journal of Signal Processing Systems 2019; 91 (2): 117-29. doi: 10.1007/s11265-018-1374-7

[15] Lahmyed R, El Ansari M, Ellahyani A. A new thermal infrared and visible spectrum images-based pedestrian detection system. Multimedia Tools and Applications 2019; 78 (12): 15861-85. doi: 10.1007/s11042-018-6974-5

[16] Bastian BT, Jiji CV. Integrated feature set using aggregate channel features and histogram of sparse codes for human detection. Multimedia Tools and Applications 2020; 79 (3): 2931-44. doi: 10.1007/s11042-019-08498-w

[17] Kumar K, Mishra RK. A heuristic SVM based pedestrian detection approach employing shape and texture descriptors. Multimedia Tools and Applications 2020; 79: 21389–21408. doi: 10.1007/s11042-020-08864-z

[18] Zhang X, Shangguan H, Ning A, Wang A, Zhang J et al. Pedestrian detection with EDGE features of color image and HOG on depth images. Automatic Control and Computer Sciences 2020; 54 (2): 168–178. doi: 10.3103/S0146411620020108

[19] Park D, Ramanan D, Fowlkes C. Multiresolution models for object detection. In: European conference on computer vision; Crete, Greece; 2010. pp. 241-254.

[20] Xu Z, Li B, Yuan Y, Dang A. Beta R-CNN: looking into pedestrian detection from another perspective. In: Advances in Neural Information Processing Systems 33; Vancouver, Canada; 2020.

[21] Wu J, Zhou C, Zhang Q, Yang M, Yuan J. Self-mimic learning for small-scale pedestrian detection. In: Proceedings of the 28th ACM International Conference on Multimedia; Seattle, WA, USA; 2020. pp. 2012-2020.

[22] Song X, Zhao K, Zhang WSCH, Guo J. Progressive refinement network for occluded pedestrian detection. In: Proceedings European Conference on Computer Vision; Glasgow, UK; 2020. p. 9.

[23] Lin C, Lu J, Wang G, Zhou J. Graininess-aware deep feature learning for robust pedestrian detection. IEEE transactions on image processing 2020; 29: 3820-3834. doi: 10.1109/TIP.2020.2966371.

[24] Karg M, Scharfenberger C. Deep learning-based pedestrian detection for automated driving: achievements and future challenges. In: Pedrycz W, Chen SM (editors). Development and Analysis of Deep Learning Architectures 1st ed. Cham, Switzerland: Springer Nature, 2020, pp. 117-143.

[25] Zhu Q, Yeh MC, Cheng KT, Avidan S. Fast human detection using a cascade of histograms of oriented gradients. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition; New York, USA; 2006. pp. 1491-1498.

[26] Zhao Y, Zhang Y, Cheng R, Wei D, Li G. An Enhanced Histogram of Oriented Gradients for Pedestrian Detection. In: Ljubo V (editor). IEEE Intelligent Transportation Systems Magazine 7th vol. Nathan Campus, Australia: IEEE, 2015, pp. 29-38.

[27] Huang J, Kumar SR, Mitra M, Zhu WJ, Zabih R. Image indexing using color correlograms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition; San Juan, USA; 1997. pp. 762-768.

[28] Chun YD, Kim NC, Jang IH. Content based image retrieval using multiresolution color and texture features. IEEE Transactions on Multimedia 2008; 10 (6): 1073-1084. doi: 10.1109/TMM.2008.2001357

[29] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence 2002; 24 (7): 971-987. doi: 10.1109/TPAMI.2002.1017623

[30] Heikkilä M, Pietikäinen M, Schmid C. Description of interest regions with center-symmetric Local Binary Patterns. In: Kalra PK, Peleg S (editors). Computer Vision, Graphics and Image Processing. Lecture Notes in Computer Science, vol 4338. Berlin, Heidelberg, Germany: Springer, 2006, pp. 58-69.

[31] Verma M, Raman B, Murala S. Local extrema co-occurrence pattern for color and texture image retrieval. Neurocomputing 2015; 165: 255-269. doi: 10.1016/j.neucom.2015.03.015

[32] Bhunia AK, Bhattacharyya A, Banerjee P, Roy PP, Murala S. A novel feature descriptor for image retrieval by combining modified color histogram and diagonally symmetric co-occurrence texture pattern. Pattern Analysis and Applications 2020; 23: 703–723. doi: 10.1007/s10044-019-00827-x

[33] Zhang B, Gao Y, Zhao S, Liu J. Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. IEEE Transactions on Image Processing 2010; 19 (2): 533-544. doi: 10.1109/TIP.2009.2035882

[34] Murala S, Maheshwari RP, Balasubramanian R. Local tetra patterns: a new feature descriptor for content based image retrieval. IEEE Transactions on Image Processing 2012; 21 (5): 2874-2886. doi: 10.1109/TIP.2012.2188809

[35] Wang Y, Zhao Y, Chen Y. Texture classification using rotation invariant models on integrated local binary pattern and zernike moments. EURASIP Journal on Advances in Signal Processing 2014; (1): 182-192. doi: 0.1186/1687-6180-2014-182

[36] Vapnik V. The Nature of Statistical Learning Theory. Berlin, Heidelberg, Germany: Springer, 1995.

[37] Mohan A, Papageorgiou C, Poggio T. Example-based object detection in images by components. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001; 23 (4): 349-361. doi: 10.1109/34.917571

[38] Munder S, Gavrila DM. An experimental study on pedestrian classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 2006; 28 (11): 1863-1868. doi: 10.1109/TPAMI.2006.217

[39] Maji S, Berg AC, Malik J. Efficient classification for additive kernel SVMs. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013; 35 (1): 66-77. doi: 10.1109/TPAMI.2012.62

[40] Mayer-Schönberger V, Cukier K. Big Data: A Revolution that will Transform How We Live, Work, and Think. London, UK: John Murray, 2013.

[41] Uddin MF, Gupta N. Seven V's of big data understanding big data to extract value. In: Zone 1 Conference of the American Society for Engineering Education; Bridgeport, CT, USA; 2014. pp. 1-5.

[42] Zhang L, Rui Y. Image search-from thousands to billions in 20 years. ACM Transactions on Multimedia Computing, Communications, and Applications 2013; 9 (1s): 1-20. doi: 10.1145/2490823

[43] Godil A, Bostelman R, Shackleford W, Hong T, Shneier M. Performance Metrics for Evaluating Object and Human Detection and Tracking Systems. Gaithersburg, MD, USA: National Institute of Standards and Technology (NIST) Interagency/Internal Report (NISTIR) - 7972, 2014.

[44] Powers DM. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies 2011; 2 (1): 37-63.

[45] Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27 (8): 861-874. doi: 10.1016/j.patrec.2005.10.010

[46] Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M. The DET Curve in Assessment of Detection Task Performance. Gaithersburg, MD, USA: National Institute of Standards and Technology (NIST), 1997.

[47] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the american statistical association 1937; 32 (200): 675-701. doi: 10.1080/01621459.1937.10503522

[48] Iman RL, Davenport JM. Approximations of the critical region of the fbietkan statistic. Communications in Statistics-Theory and Methods 1980; 9 (6): 571-595. doi: 10.1080/03610928008827904

[49] Toprak T, Belenlioglu B, Aydın B, Guzelis C, Selver MA. Conditional weighted ensemble of transferred models for camera based onboard pedestrian detection in railway driver support systems. IEEE Transactions on Vehicular Technology 2020; 69 (5): 5041-5054. doi: 10.1109/TVT.2020.2983825

[50] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine learning 2003; 51 (2): 181-207. doi: 10.1023/A:1022859003006

Abbreviations

1.	SI-HOG	Scale-invariant histogram of oriented gradients
2.	SVM	Support vector machine
3.	HOG	Histogram of oriented gradients
4.	ExHOG	Extended histogram of oriented gradients
5.	LBP	Linear binary pattern
6.	rbf	radial basis function
7.	PRNet	Progressive refinement network
8.	HSI	Hue saturation intensity
9.	BIDP	Big image data processing
10.	MR	MapReduce
11.	HDFS	Hadoop distributed file system
12.	ROC	Receiver operator characteristics
13.	FPR	false-positive rate
14.	TPR	true-positive rate
15.	DET	Detection error tradeoff