Review of Computer Engineering Research
2020 Vol. 7, No. 1, pp. 38-46
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/journal.76.2020.71.38.46

SMART FEATURE FUSION AND MODEL FOR HUMAN DETECTION

Htet Htet Lin

Assistant Lecturer, Faculty of Computer Software Computer University (BanMaw), Myanmar.
Email: htethtet.linnnnn@gmail.com Tel: +959-9-967772832

Article History
Received: 23 January 2020
Revised: 26 February 2020
Accepted: 30 March 2020
Published: 8 May 2020

Keywords
G feature
T feature
Grammar Model
Poselet Model
Smart feature fusion
Smart model fusion

ABSTRACT

Extraction of discriminate and accurate features is challenging to precise the statistical data on monitoring people. It still remains an active research due to various variations as inter class and intra class, lighting challenge, static and dynamic occlusion. To tackle this variation and occlusion issue, this paper proposes to combine the differential gradient and statistical Tamura features with joint histogram. In addition, these extracted smart features use actually to detect people by using the gradient feature descriptor and a statistical feature detector. The model fusion of human detection creates by combining two models result (Grammar model and Poselet model) with the adaptive threshold weighted non-maximum suppression algorithm. The system presents a powerful fusion insight to capture the stronger occlusion parts and several variations of the foreground people. To compare the performance with the state of the arts, the public Pascal VOC 2007 Dataset is used. The outperformed result of this work proofs our concern.

Contribution/Originality: Various feature and model was studied to detect the human. Focused on this, a fusion feature and model was contributed.

1. INTRODUCTION

Human detection is the active topic due to its widespread applications in many fields. Although many researchers have been doing this work, it still remains an active and open research problem even after research of several years in this field. There have two distinct views: sparse and dense representation. The first category is the extraction of feature from the prominent foreground regions. It does not comprise useful information but it has the detail information of small thing. The last category focuses on the foreground feature vectors to decrease the variation changes and the wrong foreground detection rate.

The texture feature extraction and analysis plays an important part in computer vision applications. It focuses many ways: on the pixel of gray value of the input image, on constructing the structural primitives from the gray pixel value, and on a model building from the neighboring gray pixel value [1]. The gradient feature is popular due to its discriminating feature properties.

Feature detection and description also plays a crucial part in people detection because the number of actual people depends on how these detected features. The quality measure of the detecting result is also based on both the accuracy of the detection features and the model representation of these features. This fact motivates to recover the various variation and occlusion point, which increase the performance detection result. The whole process of this paper is focused on each cell and overlapping blocks. The feature and model fusion are the main contribution point of this work. To investigate the generalization ability of these features, the system contributes to apply the adaptive weighted-NMS fusion algorithm.

In this paper, the effective terminology of smart feature fusion and the model has been addressed. This paper organizes as follows. Section 2 introduces the previous methods and technologies of the detection system. Section 3 presents the proposed framework to contribute the smart feature and creation model. Section 4 demonstrates the experimental results with the comparisons. The last section discussed the conclusion.

2. RELATED WORK

Human feature detection is an important and fundamental work in many surveillance monitoring system. Figure 1 demonstrates the summary of the previous human detection. Dalal and Triggs [2] studied a contrast feature set related to the edge feature direction and magnitude to know the object rate of change. They enhanced the detection result with a fixed rate scale by combining these two as self-bias. They can't solve the noisy marginal case. Zhu, et al. [3] developed a fast HOG detection system using variable-size blocks with AdaBoost. They are 70 times faster than HOG [2] but their computation cost is very high.

Mu, et al. [4] developed the semantic and fourier local binary pattern with AdaBoost. Their speed is lower than [2] but the computational cost is high. The structure of human detection is focused on the integrated motion intensity [5]. Although they introduced a filter with the threshold every 4 frames per second, their training time is very high. A two-step fast people detection framework using MSO feature classifier is proposed [6] for faster processing speed than [2]. But, they weak to handle the occlusion issue.

Wang, et al. [7] proposed a feature combination framework of local binary pattern and histogram with SVM. Although it is designed to handle occlusion issues, they can treat the partial occlusion parts. The PL-SVM framework of detection approach was proposed by joining cascaded of BO and HOG [2]. The multiple views and multiple positions of detection can be possible for higher accuracy than [2]. It operated on a cluttered background and can suppress noise. But, the time complexity is high due to noise suppression and background clustering.

3. PROPOSED APPROACH

This paper develops the framework to get the feature fusion with the model creation to solve the intra-class, inter-class and occlusion issue. The output detection quality is focused on the proposed smart feature and smart model. In the experiments, the proposed system gets the highest average precision rate than the previous works. The whole system is shown in Figure 2, where our contribution is highlighted with pink color.

In this figure, the input image from the dataset is taken as the input image of the proposed system. The proposed smart feature and model process consists of two main portions: feature fusion, and model creation with the adaptive threshold. The first feature fusion (G feature and T feature) is extracted by taking the block including cells. The joint histogram method is applied to combine these two features. In the second part, the fusion model result is developed by contributing the adaptive threshold with Grammer and Poselet model. Finally, then output result is out with the best performance.

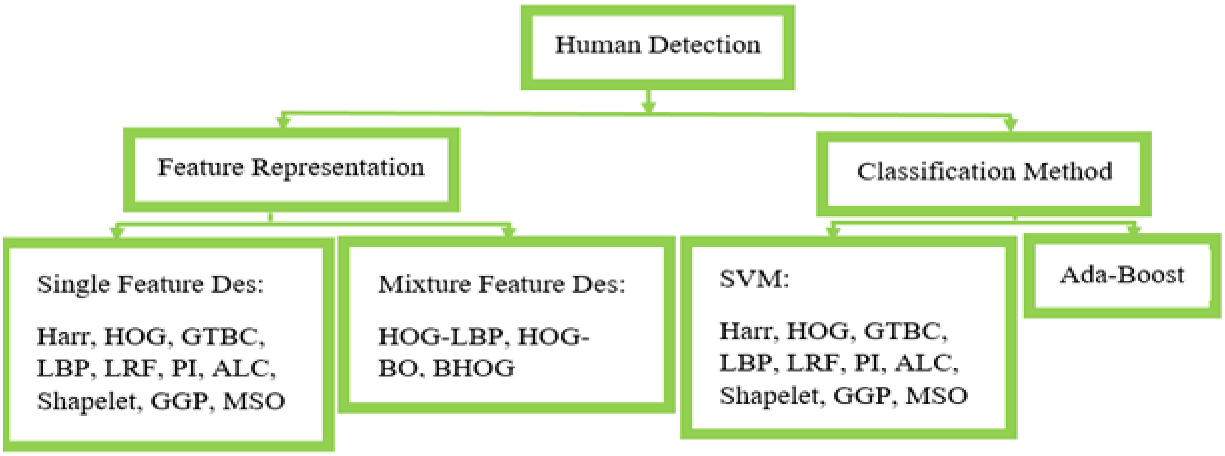

Figure-1. Topology of the person detection methods.

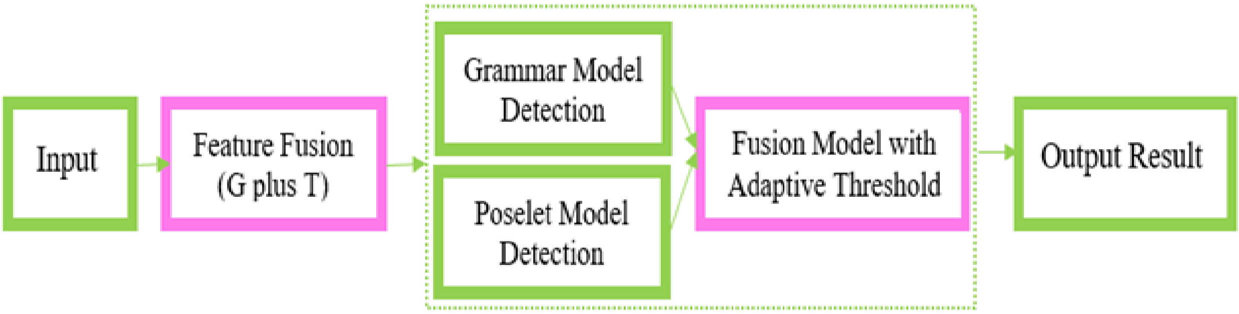

Figure-2. Proposed system overview.

a. Feature Fusion

The whole process of the first main part of the proposed system (feature fusion) is shown in Figure 3, where our contribution is highlighted with pink color. In this figure, the input is from the dataset videos. The original color channel is firstly needed to transform other hue-saturation-intensity as the following equations:

$$In = \frac{1}{3}(r+g+b) \qquad (1)$$

$$Sa = 1 - \frac{3}{I}\min(r+g+b) \qquad (2)$$

$$\theta = \cos^{-1}[\frac{\frac{1}{2}[(r-g)+(r-b)]}{\sqrt{(r+g)^2+(r-b)(g-b)}}] \qquad (3)$$

$$Hu = \begin{cases} \theta & g \leq b \\ 2\pi - \theta & g \geq b \end{cases} \qquad (4)$$

Hue and saturation values are very important to combine to get the zero order features. Intensity value is used to get the first order feature. According to the defining of k^{th}order gradients, second order gradients computes as the following equation:

$$r^* = \max_\theta \frac{\partial^2 I}{\partial u^2} \qquad (5)$$

$$\theta^* = \operatorname{argmax}_\theta \frac{\partial^2 I}{\partial u^2} \qquad (6)$$

Where I is the input intensity value and μ is the unit vector (cos ɵ, sin ɵ). The first order gradient is calculated, as like as HOG [2]. The second order gradient is very valuable for human detection. Finally, this histogram value is fused to form 'G' feature.

The proposed feature extraction process is an extension of the HOG features [2, 8, 9]. This can also see in our previous paper [9] to get the smart feature fusion. According to the psychological study of human visual perception, a new insight, Tamura is proposed to express the human texture feature, Tamura. It focuses on the probability estimation of the input image gray level by calculating coarseness, contriteness, lifelikeness, roughness, directionality, and regularity. These six features clearly showed the image size and the texture prominent element level. This extracted texture features directionality is used to form 'T' feature.

Joint histogram (JH) is used to fuse the T and G features to form smart feature fusion. It is critical in combining large images of many feature histograms, since several images have the same nature of histogram as shown in Figure 4. It uses to merge the G feature with cell based structures and T features. It reduces the space complexity by selecting the local feature set. It incorporates each entry in a joint histogram which comprises the image pixel number described by a particular feature combination value.

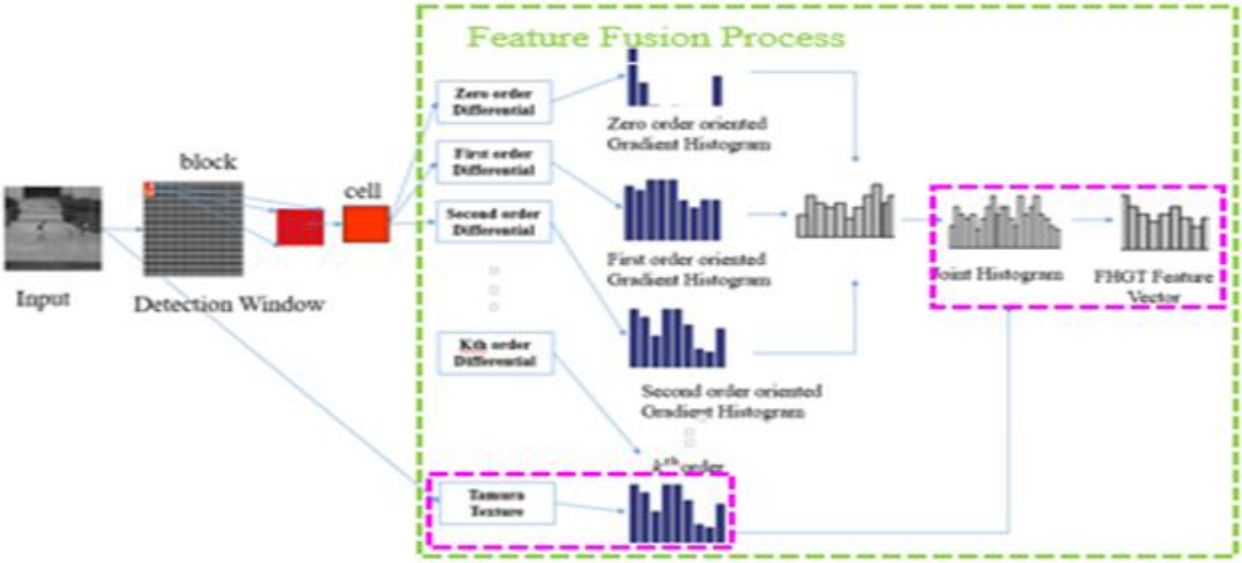

Figure-3. Proposed smart feature fusion.

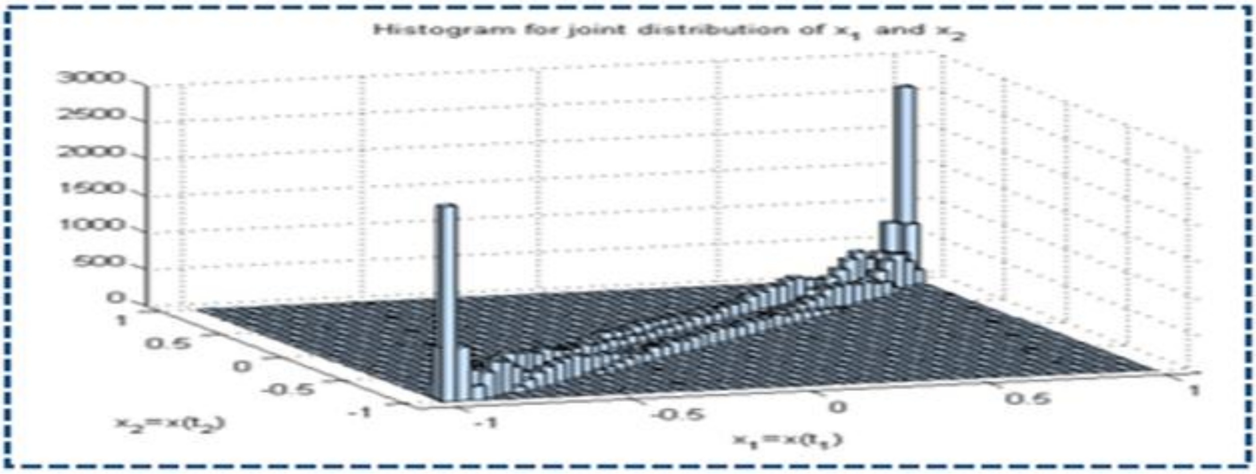

Figure-4. Joint histogram (JH).

b. Grammar and Poselet Models

Foreground objects are described by the deformation rules of object perspective. It is taken into account the moving objects respect to each other in a hierarchical deformable parts model. The structural ability of people detection is choosing between several subtypes (i.e. creating the individual model or hybrid models, etc.).

Grammar model [6] denotes the process of people detection recursively in terms of other people. Customization of each component in the people model is to detect people in a prototype visibility mode. According to this observation, it is manually developing a grammatical structure to detect people visibility.

Let N be a nonterminal symbol set, T is a terminal symbol set. Let Y be the set of possible positions for the symbol in the image. It indicates to N! (i.e. T position in one position2!). Let S be the weighted production set from the real values (R) or (r). In this paper, the structure of the grammatical model is as follows:

$$X(w_0) \xrightarrow{s} \{Y_1(w_1), \dots, Y_n(w_n)\} \qquad (7)$$

where X Ɛ N, Y_1Ɛ$N \cup T$, w_i Ɛ Ω and s Ɛ R is the score of weighted production set, denoted by r Ɛ R by s(r).

The Poselet model must combine the activations, if the same corresponding same predicted score of hypothesized, and the corresponding segmentation. It uses the symmetry to measure the consistency between two activations of i and j difference of empirical KL distribution with key points N^k as follows:

$$D_{SKL}\left(N_i^k, N_j^k\right) = D_{KL}\left(N_i^k || N_j^k\right) + D_{KL}\left(N_j^k || N_i^k\right) \qquad (8)$$

$$d_{i,j} = \frac{1}{K} \sum_k D_{KL}\left(N_j^k || N_i^k\right) \qquad (9)$$

These activations are grouped together to form human detection by using pairwise clustering based on the empirical consistency distribution. Low score and unstable activation is not able to enough to merge with one of the existing clusters will be deleted.

This is practically impossible for a single human detection model to identify all types of human body. As an example, the human grammar model cannot detect all types of human body and it only discovered some specific types compatible with the system architecture. Every model has a bias of the own theoretical limitations. There have a big difference between Grammar and Poselet models.

The Poselet model is a two-layer network based on recently defined poses. The grammatical model is focused on deformable and occluded parts. Opportunity arises from these differences. Both the Grammar model and the Poselet model are non-maximum suppressed individually before they are combined. This paper aims to improve the actually possible detection rate by the improved weighted NMS.

c. Fusion Models with Adaptive Threshold

The adaptive weighted-NMS is also described in this paper (the modification of weighted-NMS [8]). The only model of human isolation is practically impossible to detect all types of human body. For example, Grammar model, unable to detect all types of human body and their framework can't detect any particular parts. Every model is biased and embedded in their limitations. In the about section A, this paper shows the satisfactory result by combining various features see Table 1. This motivated the inspiration of different models fusion. The whole test detects and suppresses individual deviations. So, the adaptive weighted-NMS is contributed to get the best performance.

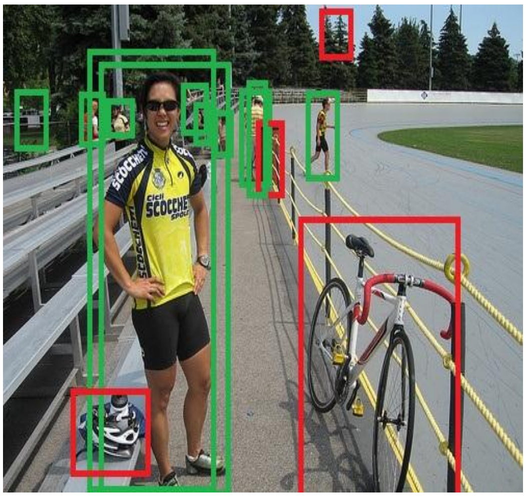

Figure-5. Example of detection result.

In Figure 5, the output of a detection model for an input image can be formulated as $(\text{pi}, \text{si})_{i=1}^{n}$, where n is the number of detections, and (pi, si) is the i-th detection. Pi is the bounding box of the position that denoted as (xmin, ymin, xmax, ymax), while the score si denotes the confidence score of the i-th detection.

However, the confidence scores from different models must have very different values or ranges of scale values. Therefore, the same estimated scores may have different levels of reliability. For example, the score range of Grammar model (-α, +α) with a value of 0 that means the semi-confidence (half), and the scope of the Poselet model (α, +α), a value of 0.5 indicates the half confidence. Therefore, the value of 0.2 exceeds the half confidence level in the Grammar model, below the level of semi-confidence in the Poselet model. In order to compare the different estimation models, we need to calibrate estimates scores in the same framework before the fusion.

So we can get the calibration score function from model II to model I as follows:

1. Record the threshold accuracy according to test the model II by curving and collecting a set of threshold points {xi}i=010, its corresponding precision value (0.0, 0.1, 0.2, 0.3, ...,0.9,1.0);
2. Perform the same operation as above for the model I;
3. Use {(Xi,Yi)}i=010 to set the carry y = f(x).

Once we get the transfer function y = f(x), we can calibrate the estimated value of model II with sʌ = f(s), where sʌ is the raw score, and s is the calibration score.

After this calibration, estimates from Model I and II will have the same range and scale. This paper introduced how to combine the detection result of the two models with higher performance. For each image we get a set of tests {(pi,si)} from model I and from the model's detection set ((Pj,Sj)) II. In here, this paper [8] assume that the estimated value of model II {sj} has been calibrated as part of Model I. It can be assumed that model I and model can be displayed with many overlap detections. Overlapping can lead to redundancy. Therefore, Non-maximum suppression method (NMS) is used to eliminate the redundant detection. Usually it is used to eliminate redundancy. But these common NMS algorithms are not suitable for this situation.

Suppose {(Pi,Si),(Pj,Sj) }are the two detections tested result from models I and II, which also have the largest overlap. The general NMS method simply removes the detection at a lower detection rate and use these scores to save the test higher score no change. Due to the large overlap between the two tests, they may correspond to the same "hypothesized object".

We believe that if a “hypothesized object” can be discovered by two different/additional models, this is more likely to be a "real object." Therefore, this is reasonable to evaluate the stored detection score instead of persistence (because this has unchanged. Based on these facts, Jiang and Ma [8] introduced the weighted NMS algorithm. This algorithm designated as a weighted NMS for merging both models. This detection is based on the low level decision that based on the modified non maximal suppression to join these two detected results of the system. In here, unlike [8] Otsu's threshold method involves iterating through all the possible thresholds due to the pi square algorithm. Although these weighted NMS can solve the issue of overlap, they still need to accurate. On the other hand, the issue of handling threshold is solved by using the weighted NMS. They don't accurate due to the stationary threshold value. This fact prompted us to modify by setting an adaptive form of threshold. We have the idea of the strong contributions that to be provided an improved NMS.

Detailed consolidation procedure based on the weighted network that displayed in Algorithm 1. First we combine detection to normalize and calibrate from these two models.

Estimate the interval corresponding to the range [0, 1]. Then, we will consider each one, Greedy, from high to low scores. If (ph,$\tilde{s_h}$) is a higher detection with the lower results (pl, sl), that is enough overlap with (ph,$\tilde{s_h}$), then (ph,$\tilde{s_l}$) will be deleted as follows:

$$\tilde{s_h} = \tilde{s_h} + w_{hl}.\tilde{s_l} \qquad (10)$$

Moreover, the P1 detection is merged into Ph and it is the weighted fraction Wh1×S1 is simultaneously absorbed by Sh. Finally, this can eliminate the redundant overlaps for the possible correct detection.

Algorithm 1 Improved weighted NMS

```
Input : Detections of model A: {($p_i$ , $s_i$ )},
Detections of model B:{($p_j$ ,$s_j$ )};// $s_j$ has already been calibrated
Output: Fused detections, i.e., the updated U
Merge$\{(p_i, s_i)\}_{i=1}^{M}$and$\{(p_j, s_j)\}_{j=1}^{N}$to a union set U : $\{(p_k, s_k)\}_{k=1}^{M+N}$;
Normalize the scores $s_k$ to interval (0, 1) with the sigmoid function
$\tilde{s_k} = {}^{1}/_{(1 + \exp\{-\propto\ .\ (s_k - \beta)\})}$;// α, β are fixed hyper-parameters
Sort the tuples $\{(p_k, \tilde{s_k})\}_{k=1}^{M+N}$ in U by descending order of $\tilde{s_k}$ ;
for h → 1 to end(U) do
    for l → h + 1 to end(U) do
        Compute overlap($p_h$, $p_l$) = $\frac{\text{area } (p_j \cap s_j)}{\text{area } (p_j \cup s_j)}$;
        if overlap($p_h$, $p_l$)>T then // threshold for overlapped detections
            $w_{hl}$ → overlap($p_h$, $p_l$); // w: decay weight for score absorption
            $\tilde{s_h} \rightarrow \tilde{s_l} + w_{hl}\ .\ \tilde{s_k}$ ;
            Delete ($p_1$, $\tilde{s}_1$) from U;
        end
    end
end
return the updated detection set U
```

The Otsu threshold technique comprises the repetition of all possible thresholds and computing pixel-level propagation metrics on each side. This means that it splits all the pixel values as two groups that arrive at the foreground or background. The goal is to discover the smoothness threshold value of the foreground and background that is minimal.

4. EVALUATION RESULT

The performance result is tested only on the person class of Pascal VOC 2007 dataset. The average precision performance matrix is used to compare with the state of the art methods. Table 1 shows the comparative results that are based on how finely extract the foreground feature is scanned. Table 2 shows the comparative results that are based on how the detection results of model are merged.

Table-1. Comparison results tested on Pascal VOC 2007.

Feature Name	Average Precision (%)
HSC Feature [10]	44.14
HOG Feature [2]	45.8
HOG III Feature [8]	51.3
G Feature without same dimensions [9]	50.1
G Feature [9]	51.3
T Feature [9]	38.0
Smart Feature (Fusion of G and T by JH)	52.1

Table-2. Different models results tested on Pascal VOC.

Model Name	Average Precision (%)
Grammar Model with Smart Feature	46.8
Poselet Model with Smart Feature	49.6
Smart Model with Smart Feature	55.3

5. CONCLUSION

A robust human detection system by fusion the feature and model was designed to solve the occlusion and variation issues. People contours and gradients are firstly extracted from the salient foreground by developing the smart feature. The model fusion not only combines multiple detections at nearby places, but also includes of the detection, but also allows the reliable detection of object occlusion that appear at very different scales. The experimental results are outperformed than the previous approaches.

Funding: This study received no specific financial support.
Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.
Acknowledgement: The author wishes to thank her parents, elder brother and relatives whose inspirations make their work hard and who are always willing to give the authors all moral support.

REFERENCES

[1] P. P. Reboucas Filho, E. D. S. Reboucas, L. B. Marinho, R. M. Sarmento, J. M. R. Tavares, and V. H. C. De Albuquerque, "Analysis of human tissue densities: A new approach to extract features from medical images," *Pattern Recognition Letters*, vol. 94, pp. 211-218, 2017.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 886-893.

[3] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *In Proceeding IEEE International Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1491–1498.

[4] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *In Proceeding IEEE International Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[5] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision* vol. 63, pp. 153–161, 2005.

[6] Q. Ye, J. Jiao, and B. Zhang, "Fast pedestrian detection with multi-scale orientation features and two-stage classifiers," in *In Proceeding IEEE 17th International Conference on Image Processing*, 2010, pp. 881–884.

[7] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *In Proceeding IEEE International Conference Computer Vision*, 2009, pp. 32–39.

[8] Y. Jiang and J. Ma, "Combination features and models for human detection," in *In Proceeding IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] H. H. Lin and K. T. Win, "Person detection based on fusion histogram of gradients with texture (FHGT) local features", in Researchscript," *International Journal of Research in Computer Scientific (IJRCS),* vol. 5, pp. 1 – 4, 2018.

[10] X. Ren and D. Ramanan, "Histograms of sparse codes for object detection," in *In Proceeding IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.