IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 14, 2021, accepted February 4, 2021, date of publication March 1, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063028

Vision-Based Human Detection Techniques: A Descriptive Review

SHAHRIAR SHAKIR SUMIT[1], DAYANG ROHAYA AWANG RAMBLI[1], AND SEYEDALI MIRJALILI[2,3], (Senior Member, IEEE)

[1]Department of Computer and Information Sciences, Universiti Teknologi PETRONAS (UTP), Seri Iskandar 32610, Malaysia
[2]Centre for Artificial Intelligence Research and Optimization, Torrens University Australia, Fortitude Valley, QLD 4006, Australia
[3]Yonsei Frontier Laboratory, Yonsei University, Seoul 03722, South Korea

Corresponding author: Shahriar Shakir Sumit (shahriar9121@gmail.com)

This work was supported by the Yayasan UTP through the Cost Centre under Grant 015LCO-001.

ABSTRACT Cameras are being used everywhere for the safety and security of citizens in different countries. Using a machine to detect humans in a photo or a video frame is a very complicated and challenging task. Various techniques have been developed for this purpose, which mainly rely on Artificial Intelligence. This article aims to provide a comprehensive review and analysis of the literatures from a descriptive perspective, which is its main differentiator from the existing survey papers in this area. Firstly, the vision-based human detection techniques and classifiers are elucidated in conjunction with the variants of feature extraction techniques. Secondly, various pros and cons of such techniques are discussed. Then, an investigation has been conducted and reported based on the state-of-the-art human detection descriptors (e.g. Log-Average Miss Rate and accuracy). Although techniques such as Viola-Jones and Speeded-Up Robust Features can detect objects in real-time and overcome Scale-Invariant Feature Transform (SIFT) limitations, they are still sensitive to illuminated conditions. Other techniques such as SIFT, Bag of Words, Orthogonal Moments, and Histogram of oriented Gradients provide other interesting benefits which include insensitivity to occlusion and clutters, simplicity, low-order element construction and invariance to illuminated conditions; nevertheless, they are computationally expensive and sensitive to image rotation. A meticulous review along similar lines revealed that the Deformable Part-based Model performs relatively better due to its ability to deal with particular pose variations and multiple views, occlusion handling (partial) and is application-free while its counterparts focus on only a single aspect. This article highlights and provides a brief description of each available data-sets for human detection research. Various use-cases of human detection systems are also elaborated. Finally, various conclusions are derived based on the conducted review followed by recommendations for future directions and possibilities to further improve the speed and accuracy of human detection systems.

INDEX TERMS Computer vision, object detection, human detection, feature extraction techniques, classifiers.

I. INTRODUCTION

A novel coronavirus (COVID-19) pandemic [1], [2] affecting the respiratory portion of the human system is currently ongoing [3] causing a high degree of mortality and morbidity globally [4]. More than 2.5 million people lost their lives and 113 million people are infected by this virus across the globe [5] as of February 2021. There is no proper treatment yet to cure this disease. This virus is contagious and can be transmitted from one infected person to another person. According to the World Health Organization (WHO) information, its spread can be reduced by maintaining social distancing, wearing a face mask and washing hands [6]. In order to reduce the spread of this virus, human detection can play a very important role, for example, by checking whether people wear facial mask correctly or follow social distancing instructions.

With the growth of modern technology, human detection can play a major role in avoiding or minimizing accidents or death due to natural calamities. Every year, approximately 60,000 people lost their lives due to natural disasters, which is, 0.1 percent of the world total deaths [7]. In Australia,

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed Farouk.

for instance, more than 28 people were killed and over 3,000 houses were destroyed by bushfire [8]. At least 47 people were reported dead in volcanic eruption in New Zealand [9]. Using Unmanned Aerial Vehicles (UAV) [10], imagery human can be detected and therefore, accidents or deaths can be avoided in these areas.

Human detection can play a key role in e-health applications, for instance, to detect the fall of elderly people. The population of the elderly is increasing worldwide. By 2050, the number of old people will increase to more than 20 percent in the United States and more than 30 percent in the European countries and China mainland [11] and around the world, the elderly population will reach 2 billion [12]. Many studies have cited the fall incident as a serious issue for aging people because the majority of these elderly live by themselves [13].

Annually, an estimated 1.35 million people lost their lives and 20 to 50 million people around the world experience non-fatal complications because of traffic accidents [14]. In the United States, more than 5,000 pedestrians lost their lives due to road accidents and around 130,000 pedestrians needed medical treatments with non-fatal complications in 2015 [15]. Traffic death ratio can be minimized by using autonomous car techniques and devices [16], [17] which are capable of interacting with people [18].

With the continuous increase in the crime rate and public concern due to terrorist activities for the past few decades, public security is an inevitable consequence and human detection systems, can be used to monitor, manage and enforce law in public areas. Every year around 21,000 people lost their lives due to terrorism. 0.05 percent of the total worldwide deaths in 2017 is caused by terrorism [19]. The need to install a large number of human detection systems has increased significantly in public areas after the terrorist attacks in New York [20], London [21] and some other cities in the world. These happenings are serious enough that robust detection systems need to be designed and deployed in such scenarios. The role of human detection system is considered to be a promising solution for ensuring public safety; hence, it has become one of the important areas of research in the field of computer vision.

Accurate identification of particular objects is essential in order to perceive and understand the specific details of images. Object detection, which includes identifying the existence of a certain object in a picture and figuring out its location, is considered as a key problem in the field of computer vision. Moreover, image understanding includes not only detecting objects in a particular scene but also categorizing the detected objects in their respective classes [22]. Visual object detection is primarily concerned with the classification of object categories [22]. Figure 1 shows an example of the basic work flow of a machine learning paradigm for object detection. For instance, we want to classify three different objects: a cat, a human and a horse. In order to train a machine learning model, the first task is to collect training images with label data. The next task is to extract the features and add it to the classification model.

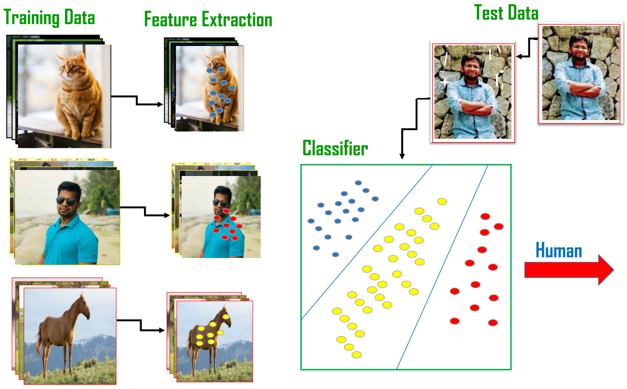

FIGURE 1. Example of machine learning work flow for object classification across three different classes (cat, human and horse).

Features [23] can be represented by any object properties such as colours, edges, corners, blobs or regions and ridges. The success of the training process depends on the feature extraction, classifier selection and training step (i.e. the iteration process). Among these steps, feature extraction is very important to get the desired outcome as it enhances the trained models' accuracy from the input data by removing unwanted or unnecessary features. Feature extraction minimizes the data dimensionality by extracting the redundant data, thus improves the inference and training speed. Ideally, we want the features to be invariant to different lighting conditions, and should have the ability to handle changes in scale or rotation. Upon extracting all possible features from the images, these features are added to the training model. Then, they are fed to a suitable type of classifier, depending on the speed and accuracy. Nearest Neighbour classifier (NN), Decision Tree and Support Vector Machine (SVM) are commonly used as classifier models. Once the training model is set, then for a given test image, it is possible to extract the same features from the test image and as a result, to predict the correct category class from those features by using the trained model.

Human can easily detect objects in images or videos without difficulties. They can understand emotional status, relationship among people and count the number of human occurrences in the pictures. Computer vision is expected to give technological assistance for this human capability. The computer vision's aim is to make a computer realize a picture or a video with the help of the eyes of a digital camera. Computer vision is a research field that entails different techniques to acquire, analyze, process and understand pictures [24].

Human detection is one of the core tasks in computer vision. It is challenging to detect human in images due to various factors which include illumination conditions [25], background clutter [25] and occlusion (partially or fully detected) [26]. Early methods had failed to detect humans in a real world environment [27], took more time [28] and produced less accurate results [29] due to distance [30] and appearance changes [31]. These factors make it extremely problematic to find a universal representation for the human object. Despite these shortcomings, human detection is being used in many applications (for details see the application part).

Review or survey articles can be categorized into two major groups known as Review and Comparison. These two groups can be further divided into four divisions: Brief Reviews, First Level Evaluation, Deep Evaluation and Descriptive Reviews [32]–[34].

We have investigated the existing vision-based human detection surveys, in which the authors reviewed human detection mostly for specific applications. These examples include detecting pedestrians for safety in "driving assistance" method [35]–[39], human activity recognition [40] and "human motion" analysis method [41]–[43].

Zhou *et al.* have given a brief review on the human detection and tracking techniques from a diagnostic and clinical perspective. The authors focused on the variation of non-visual tracking methods with sensing techniques (inertial, magnetic and other sensors) and visual tracking methods (marker-free or marker-based) [44]. Gandhi *et al.* have written another comprehensive review on human detection focusing on mainly "Pedestrian Protection Systems (PPSs)". Furthermore, the authors included probabilistic models to analyze the human behavior with the purpose of avoiding collisions between vehicles and human [35].

Schiele *et al.* have written a deep evaluation review on human detection in which the authors mentioned the improved performance combining laser-based and visual human detection techniques. Moreover, the authors provided a potential research guideline for visual human detection [45]. Geronimo *et al.* have given a comprehensive deep evaluation review for human detection. The authors detected human for "Advanced driver assistance systems (ADASs)" focusing especially on "pedestrian protection systems (PPSs)" while mentioning future guidelines and challenges in the field [37].

Paul *et al.* have provided a first level evaluation review on human detection. The authors focused on the conventional approaches: spatio-temporal filtering, optical flow and background subtraction methods to detect the human from the surveillance videos. For classification purpose, the authors used motion-based, texture-based or shape-based features [46]. Nguyen *et al.* have written an inclusive first level evaluation survey for human detection. The authors discussed some human detection problems such as real-time issues and partial occlusion. Besides, the authors mentioned the failures and successes of the existing methods [47]. Brunetti *et al.* have provided another first level evaluation review for human detection, particularly focusing on pedestrians. The authors reviewed vision-based techniques with deep learning, especially CNN (Convolutional Neural Networks) techniques for pedestrian detection and tracking [48].

We have observed only one short descriptive review so far on human detection [49]. Nevertheless, the article did not satisfy the descriptive review requirements. The authors described several descriptors while ignoring some revolutionary methods. Besides, the authors did not mention the pros and cons of the described methods. This study is different from the existing review articles aforementioned in the following areas:

- The key objective of this article is to offer a comprehensive review on human detection utilizing various machine learning techniques.
- Vision-based on revolutionary methods with their variants are described in this article while mentioning the successes and failures of various methods/models.
- Available human descriptors are investigated and reported based on their log-average Miss Rate (MR) %) and accuracy.
- Human detection applications are discussed and available data-sets are listed with brief description.
- A comprehensive guideline is provided to improve these descriptors, particularly for speed and accuracy.
- Open problems and future directions are elaborated to provide a foresight of the possible gaps that also deserve additional attention and which can later be addressed as part of the future work in the area of human detection.

In general, this review article lays the necessary ground for those who are keen to explore the area of human detection based on computer vision.

Before heading for the full study, it is necessary to address the outline of the present work. The contents are as follow: (section II) explains the human detection methods and their variants with their successes and challenges, (section III) illustrates the classifiers with their pros and cons, human detection applications are described with advantages and limitations (in section IV), available human detection datasets with brief description are discussed (in section V), state-of-the-art human detection methods' results are reported (in section VI), (in section VII) some suggestions to improve the existing descriptors are provided and open issues problems and future direction are given, and finally, section VIII concludes the article.

II. FEATURE EXTRACTION TECHNIQUES

In the past decade, researchers have been paying much attention to design better hand crafted features to enhance the object detection accuracy and robustness. A large number of standard features have been offered by computer vision community for object detection, for example "scale-Invariant Feature Transform (SIFT)" [50], "Speeded Up Robust Features (SURF)" [51] and "Histogram of oriented Gradients (HoG)" [52]. Figure 2 provides a taxonomic view of various feature extraction techniques for human detection.

In the subsequent sections, we summarize some of the most popular feature extraction methods.

A. VIOLA-JONES MODEL

The Viola-Jones [53] face detection algorithm is the first framework for robust face detector – proposed in 2001 by

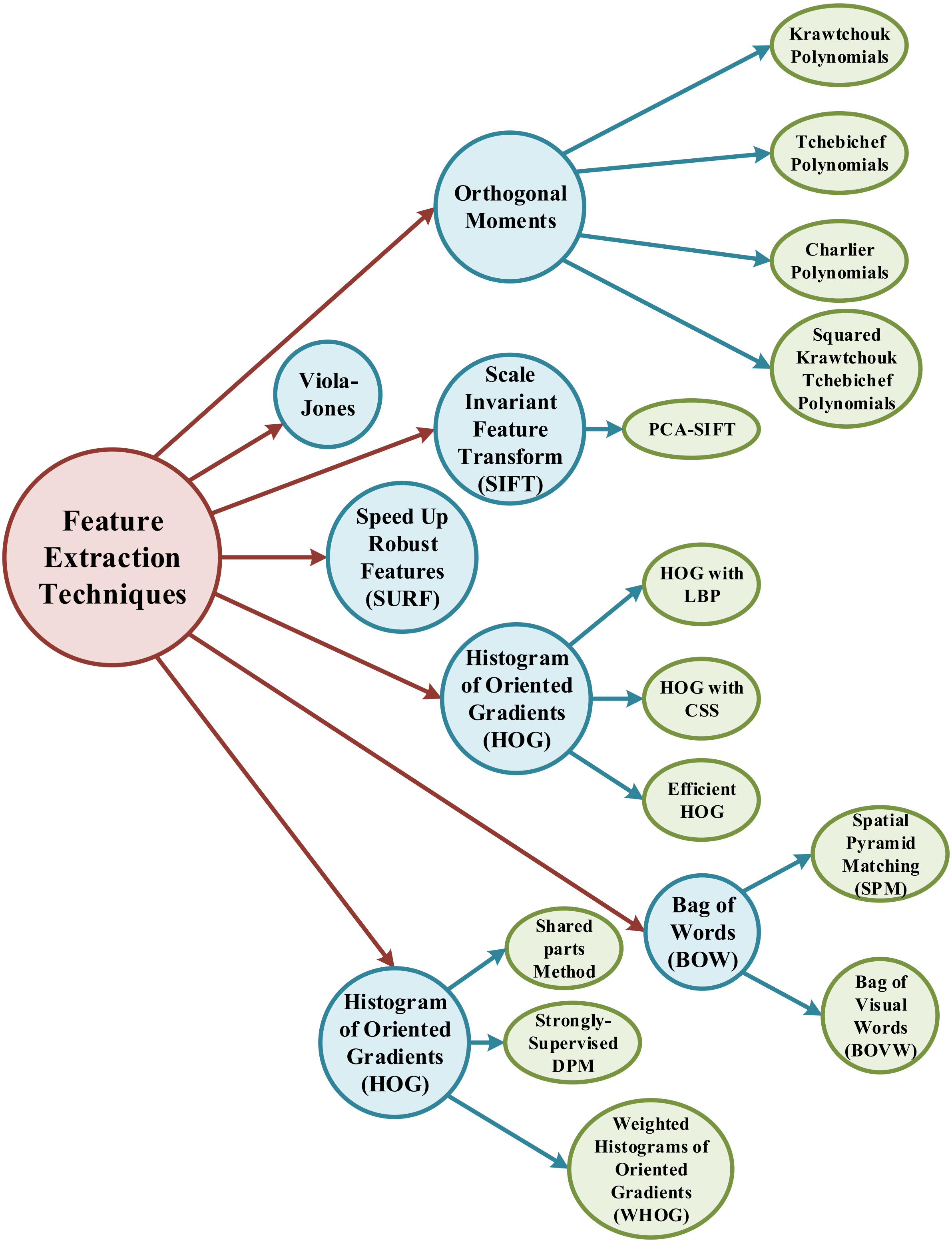
Feature Extraction Techniques
Orthogonal Moments
Krawtchouk Polynomials
Tchebichef Polynomials
Charlier Polynomials
Squared Krawtchouk Tchebichef Polynomials
Viola-Jones
Scale Invariant Feature Transform (SIFT)
PCA-SIFT
Speed Up Robust Features (SURF)
Histogram of Oriented Gradients (HOG)
HOG with LBP
HOG with CSS
Efficient HOG
Bag of Words (BOW)
Spatial Pyramid Matching (SPM)
Bag of Visual Words (BOVW)
Histogram of Oriented Gradients (HOG)
Shared parts Method
Strongly-Supervised DPM
Weighted Histograms of Oriented Gradients (WHOG)

FIGURE 2. Taxonomy of various feature extraction techniques for human detection.

Viola and Jones [53]. The system uses Haar features and is able to perform real time detection of faces. Even though, the training of the system is slow, the detection is fast.

During the training phase, small sub-windows of different shapes (shown in Figure 3(a)) are applied across the bounding boxes in the training images. The features are computed by

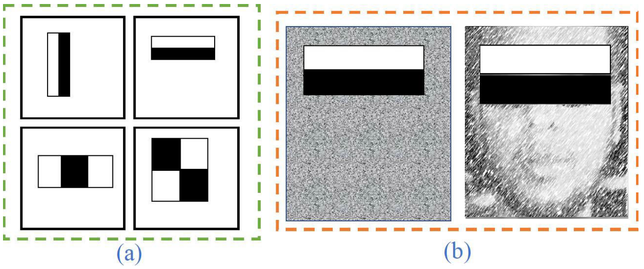

FIGURE 3. (a) Shows different Haar features. (b) Example of Haar feature that produces no signal when applied on a random distribution of intensity values (on the left) but the same Haar feature, when placed over forehead and eyes (right side), produces some signal.

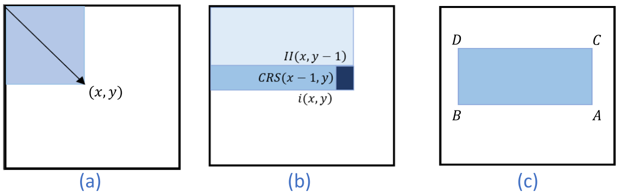

FIGURE 4. Constructing an integral image.

the following Equation (1):

$$I = \sum I_{wa} - \sum I_{ba} \tag{1}$$

where I_{wa} and I_{ba} are the intensities of white and black areas respectively.

The result of applying such a sub-window will result in a scalar value for a gray scale image. In case of color image, it will be a vector of three elements – based on the three color channels. Figure 3b shows an example of applying such sub-window to different images. The left image of Figure 3b will produce no signal or zero when a particular sub-window is selected for feature computation; nonetheless, with the same sub-window, when placed right across the forehead and eyes of a human face (right image in Figure 3b) will produce some signal. The signal will not be very strong as the system is not robust i.e. they will form a weak learner. The authors came up with the idea that by using several of these weak learners – which are able to produce signals at different locations of a face with at least an accuracy of more than 50% – a robust face detector can be built.

One major problem with this approach is that summing all the values under the sub-window and also with many variations of sub-windows can be computationally very expensive. To solve this issue, the authors proposed a concept of integral images for faster computation. This simple concept of integral image is shown in Figure 4c. In an integral image, the value of a particular pixel (x, y) is represented as the sum of all the pixels under the region shown in blue rectangle in Figure 4c i.e. integrating all the pixels in a 2D space.

$$CRS(x, y) = CRS(x-1, y) + i(x, y) \tag{2}$$

$$II(x, y) = II(x, y-1) + CRS(x, y) \tag{3}$$

Equation (2) and (3) show the procedure on how to compute the integral values from an image. Where i represents the original image, CRS denotes the cumulative row sum, and II denotes the integral image. Thus only 3 operations ((i) computing the integral value at location $(x, y-1)$), (ii) cumulative sum of the current row upto the point $(x-1, y)$ and (iii) finally adding them together) are required to compute the value of the integral image at any location and it will be $3n$ where n is the pixels in an image. So it is a linear operation in time.

$$sum = A - B - C + D \tag{4}$$

where A, B, C, D indicate the coordinates of rectangle $ABCD$. Now given any rectangle (see Figure 4e) over the image, the sum of original image values within the rectangle can be computed by using the Equation (4). Thus to obtain the sum of all the pixels in the original image, we only need to perform three addition/subtraction operations. Subtracting B and C from A in the integral image will subtract point D twice, thus in the formula, D is added again.

In the next stage, a classifier is built with the help of boosting. Boosting is a learning algorithm that takes all the weak learners and produces a accurate ensemble classifier out of those weak learners. Before we train the system, data should be labeled with the bounding boxes of a human face which will be serving as positive samples and any regions without the faces can be used as negative samples. Prior to the training, some fixed weights are assigned for the positive and negative samples. During each iteration of the training, one weak learner is selected that gives the best performance and raise the weights of the positive samples that were misclassified by the selected weak learner. Finally, the classifier is computed as a linear combination of all the weak learners where the weights of each learner is proportional to its accuracy. As the boosting iteration increases, every time the system has to out perform the previous round of the boosting. Hence, the system gets better with time as more and more features are added. Among various boosting schemes, AdaBoost is the most commonly used. In this method, a series of classifiers are trained in a cascading manner. The combination of complex classifiers in a cascade structure increases the speed of the detector dramatically by focusing attention on promising regions of the image.

1) SUCCESS

Integral image is used to calculate rectangle features for object detection that maximizes detection accuracy while minimizing computing time. Even though the train time is higher, it can perform real time face detection after training. Viola and Jones can detect walking human on low resolution images under difficult conditions such as rain and snow by implementing of a state of the art pedestrian detection system [54].

2) CHALLENGE(S)

This method lacks all the fine details of the image as it only relies on coarse intensity difference. It also does not account for any texture or shape information. The system is very much

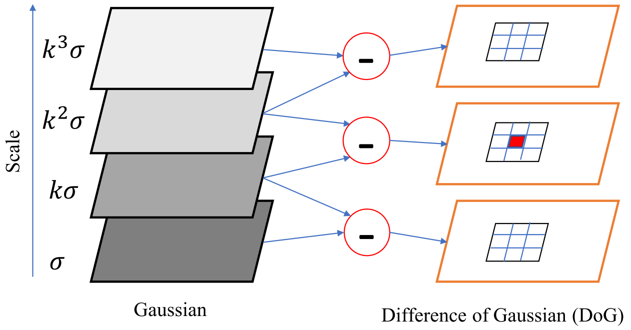

FIGURE 5. Schematic representation of constructing scale space images using different values of sigma. Edge maps at different scales are generated by taking pairwise difference of Gaussian scale output to produce the Laplacian of Gaussian (LoG).

dependent on sensitivity of the intensity values; any changes on lighting condition will result in failure in the detection phase. This is also not suitable for general object detection [55], [56].

Table 1 illustrates a brief evaluation of Viola-Jones, SIFT, PCA-SIFT and SURF methods.

B. SCALE INVARIANT FEATURE TRANSFORM (SIFT)

Scale Invariant Feature Transform (SIFT) [62] is one of the most influential methods in the computer vision community around in the year 2004. It is commonly used for recognizing objects, once the system is trained with few example images of the same object. It is an interest point detector (also known as key point extractor) that addressed the problem of matching features by providing local descriptors for the key points which is capable of coping with changes in scale and rotation.

While detecting edges using techniques like Canny edge detector [71] or by using Laplacain of Gaussian, the primary challenge is to select the value of sigma (σ) i.e. what should be the width of the mask and there is no easy answer for that. To overcome this problem, the author used many values of sigma by creating a scale space of any given image. The scale space is created by using Gaussian function with progressively increasing sigma (σ) value, which produces blurred images at different scales (see Figure 5). The scale space of an image is defined as a function, $L(x, y, \sigma)$, where L represents the level of the scale space and is produced from the convolution of a variable-scale Gaussian, with an input image, $I(x, y)$ by using the equation (5).

$$L(x, y, \sigma) = G(x, y, \sigma) \circledast I(x, y) \tag{5}$$

where $\circledast$ is the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-}(x^2 + y^2)/2\sigma^2 \tag{6}$$

From the heat equation for Gaussian, we can take the derivative of Gaussian (G) with respect to sigma (σ), it should be equal to the Laplacian of Gaussian (LoG) ($\nabla^2 G$), multiplied by sigma (σ).

$$\frac{\partial G}{\partial \sigma} = \sigma \nabla^2 G \tag{7}$$

Here Gaussian (G) is function of three variables (x, y, and σ). From the approximation of derivative, we can compute the rate of change of G with respect to σ by subtracting the values of Gaussian at some point σ with another value of Gaussian at another point $k\sigma$ and then by dividing the difference ($k\sigma - \sigma$) – as shown in the following equation (8).

$$\sigma \nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x, y, k\sigma) - G(x, y, \sigma)}{k\sigma - \sigma} \tag{8}$$

The equation (8) can be rewritten in the form of equation (9).

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k - 1)\sigma^2 \nabla^2 G \tag{9}$$

Equation (9) shows that we can approximate the Laplacian of Gaussian by taking the difference of Gaussian filter at σ and $k\sigma$ - which is an efficient way to compute LoG (see Figure 5). Laplacian of Gaussian (LoG) is then employed to identify the potential interest points.

However, the obvious question is how do we combine all the edge maps to select the appropriate interest or key points? According to this algorithm, if we want to check a point whether it is a interest or key point or not (see the red pixel in Figure 5), we need to scan 3 × 3 neighborhood of the same scale as well as one scale above it and also one scale below it. The pixel is only considered to be an interest or key point if it is a local extrema i.e. either the value is the largest or the lowest out of all 26 points from the 3 × 3 neighbourhood including the adjacent levels (Figure 5). Note that for 3 × 3 neighbourhood, each level will have 9 points, all together - including the adjacent levels (the level below and the level above) - it is 27 points and the central point (shown in red color (Figure 5)) is compared with rest of the 26 points. Ultimately, all these key points are finding blobs. The size of the blob is determined by the value of σ.

Once the key points are extracted, the SIFT algorithm also provides a way to describe the SIFT key points (also known as features) which are used to compare while recognizing the same object. To generate the descriptor around each key point, a histogram of local gradients are computed for the selected σ value i.e. the scale at which the blob was detected. This removes the rotational uncertainty while matching the features to another image. A threshold is applied to discard the outliers from the key points based on minimum contrast. Additionally, another threshold is applied based on the ratio on principle curvatures to ensure the reliable orientation. Now for each key point, in a 16 × 16 neighborhood, all the gradients are quantized into 8 different directions and formed a histogram. Then it splits that window into sixteen 4 × 4 windows. From each 4 × 4 window it generates a histogram of 8 bins. In this case, there will be 128 elements in the final feature descriptor, which can be normalized to unit length to handle illumination differences.

Success: SIFT features are not dependent and do not change when an image is scaled or rotated. The partially invariant to lighting and the 3d viewpoint of a camera though. SIFT features are local features, which make them robust

TABLE 1. A brief summary of Viola-Jones, SIFT, PCA-SIFT and SURF techniques.

methods	application	advantages	limitations
Viola-Jones	object detection, face detection [53], pedestrian detection [57]	• takes less computational time while maintaining high accuracy in real-time [53] • can detect object under complex situation (i.e. rain , snow) [53] • can successfully detect pedestrian in low resolution [54]	• lacks the entire image fine details [55], [56] • texture or shape information are ignored [55], [56] • sensitive to lighting condition [55], [56] • unsuitable for general object detection [55], [56]
SIFT	object recognition, face recognition [58], gesture recognition [59], video tracking [60], motion tracking [61]	• robust to occlusion, clutter and noise [62] • distinctive features [62] • performance is close to real-time [63] • flexible to extend with other features [64]	• poor performance with lighting changes and blur [65] • computationally expensive [65]
PCA-SIFT	object recognition, image retrieval [66], image analysis [65]	• reduces the dimensionality of the SIFT descriptors [66] • improves the matching accuracy and speed in real-world environment [66]	• sensitive to viewpoint change [65] • color information is ignored [67]
SURF	object recognition [51], [68], face detection [69], image registration [66], object classification [70]	• takes less time for computation and feature matching [51] • improves the robustness of feature extraction [51]	• struggles under scale, rotation and blur invariance [65] • sensitive to viewpoint change, illumination condition [65]

to occlusion, clutter and noise. The features are highly distinctive, individual features which can be matched to a large database of objects, thus providing a basis for robust object recognition, motion tracking [62]. It is more accurate and better descriptor compared to other older descriptors [72] at that time.

Challenges: SIFT does not use global information and relies heavily on local information. Generally speaking, it does not perform well with lighting changes and blur. SIFT is based on the Histogram of Gradients. The gradients of each Pixel in the patch need to be computed and these computations cost time [65].

1) PCA-SIFT

PCA-SIFT [66] is a variation of SIFT which produces alternative representation. Principal Components Analysis (PCA) is employed to the normalized gradient patch in lieu of SIFT's smoothed weighted histograms. The only change in PCA-SIFT is the construction of key point descriptors. PCA-SIFT pre-computes an eigen-space for local gradient patches of size 41 × 41. Next, it calculates 39 × 39 horizontal and vertical gradients, and results in a vector of size 3042 (normalized image gradient vector). Later on PCA-SIFT multiplies this vector to get a compact feature vector using the stored eigenspace $n \times 3042$ projection matrix. To determine whether two vectors in two images correspond, Euclidean distance is used. This results in a PCA-SIFT descriptor of size n, where $n = 20$.

a: SUCCESS

PCA-SIFT produces alternative representation which improves upon the local image descriptor than SIFT. It reduces the dimensionality of the vector using PCA (from 3042 to 36). PCA-SIFT shows remarkable progress in matching the accuracy and speed for regulated and real-world environments [66].

b: CHALLENGES

PCA-SIFT is sensitive to viewpoint change. PCA-SIFT performs very well when the viewpoint angle is smaller than 30 degrees but its performance tends to worsen when the viewpoint angle becomes greater than 30 degrees [65].

C. SPEEDED UP ROBUST FEATURES (SURF)

SURF [51] is a Hessian based scale and rotation invariant interest point detector and descriptor for object detection. The SURF approach is a robust and swift technique for local representation of invariant similarities and image comparison. The SURF approach keeps a key interest in its operators' quick computation employing box filters. SURF contains two steps: a) Feature Extraction and b) Feature Description.

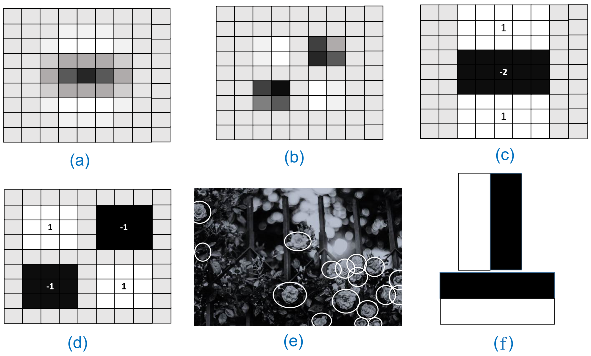

FIGURE 6. (a)-(d): The Gaussian second order partial derivative (discretised and cropped) in y-direction(G_{yy}) and xy- (G_{xy}), accordingly, (e) shows interest points detection & (f) shows the Haar-wavelet that are employed for orientation assignment [51].

The approach adopts a simple "Hessian matrix" approximation for key point detection in the feature extraction phase because of its excellent results in accuracy and calculation time. Surf depends on the Hessian matrix's determinant to select the position and scale. For any pixel of the image P at point x, the Hessian is defined by

$$H(g(x,y)) = \begin{bmatrix} \dfrac{\partial^2 g}{\partial x^2} & \dfrac{\partial^2 g}{\partial x \partial y} \\ \dfrac{\partial^2 g}{\partial x \partial y} & \dfrac{\partial^2 g}{\partial y^2} \end{bmatrix} \tag{10}$$

Gaussian Kernel filters the image, in ordered tuple $X = (x, y)$ in a image P, the "Hessian Matrix" $H(x, \phi)$ in x at scale ϕ is expressed by

$$H(x,\phi) = \begin{bmatrix} G_{xx}(x,\phi) & G_{xy}(x,\phi) \\ G_{xy}(x,\phi) & G_{yy}(x,\phi) \end{bmatrix} \tag{11}$$

where $G_{xx}(x, \phi)$ is known as the convolution of the Gaussian second order partial derivatives and in the same process $G_{xy}(x, \phi)$ and $G_{yy}(x, \phi)$ are defined.

Gaussians are most advantageous for scale-space interpretation, but the Gaussian needs to be discretised and cropped in practice. This causes the repeatability loss which is subject to image rotation in the region of odd $\pi/4$ multiples. This shortcoming generally holds for Hessian dependent detector. At first, Gaussian Kernel convolution needs to be applied and then partial second-order derivatives compute the Hessian matrix determinant. For example, the 9×9 box filters (shown in Figure 6a,b,c,d) are approximated by Gaussian second order partial derivatives with $\phi = 1.2$ and is represented at the lowest scale (i.e. highest spatial resolution). These approximations are denoted by M_{xx}, M_{yy}, and M_{xy}. The weights applied to the rectangular regions are kept simple in the expression for computational efficiency and also required re-balancing the relative weights for the Hessian's determinant defined as

$$\frac{|G(1.2)|\,|M(9)|}{|G(1.2)|\,|M(9)|} = 0.912\ldots \simeq 0.9 \tag{12}$$

where $||x||_F$ is the Frobenius norm. This results

$$|(H_{approx})| = M_{xx}M_{yy} - (0.9M_{xy})^2. \tag{13}$$

SURF converts the image sequences into integral image. Integral image can calculate the pixel intensities in square region quickly. They permit for the very quick execution of box type convolution filters. The element of an integral image $P_{\sum(x)}$ at a position $X = (x, y)$ describes the sum of all pixels in the input image P of a rectangular region formed by the point x and the origin. This can be evaluated by

$$P_{\sum}(x) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} P(i,j). \tag{14}$$

When the value of $P_{\sum}$ is occurred, only four additions are required for the sum of the intensities over any upright, rectangular area, independent of its size. As box filters and integral images are used, it is not necessary to employ the same filter to the result of a formerly filtered layer. Any size of filters can be applied parallelly on the original image with the same speed.

The scale space is then analyzed by up-scaling the filter size rather than iteratively decreasing the size of the image. Then the filter feed-backs are normalized subject to the mask size. It ensures a constant "Frobenius norm" for every filter size. Scale spaces are normally enforced as image pyramids. The images are continually smoothed with a Gaussian and afterwards sub-sampled to obtain a maximum advanced level in the pyramid. The result of the above 9×9 filter is known as the initial scale layer, where scale $s = 1.2$ and $\phi = 1.2$. Moreover, as the "Frobenius norm" remains constant for all filters, they are already scale normalized. A non-maximum concealment is employed in a $3 \times 3 \times 3$ neighborhood to locate the key points in the picture and on scales.

After that, the maxima of the Hessian matrix determinant is interpolated into the scale and image space in the feature description phase. Scale space interpolation is important particularly in this case, as the difference in scale between the first layers of every octave is relatively large. (Figure 6e) shows an example of the detected interest points in rose garden using "Fast-Hessian" detector. Haar wavelets (shown in Figure 6g) of size $4s$ are computed for the orientation assignment in x and y directions. Haar-wavelet responses are calculated for pixels that are located within a radius of $6s$ around the interest point. The dominant orientation is evaluated by sum of vertical and horizontal responses. The descriptor is calculated using Haar wavelets in square $20s$ size area centred at the interest point and oriented along the dominant direction.

1) SUCCESS

Integral images are used for speed in SURF. Moreover, SURF has minimized the time for computing and matching features and also improved the robustness with just 64 dimensions simultaneously [51].

2) CHALLENGES

In terms of matching correct key-points under scale and rotation invariance, SURF algorithm does not provide a satisfactory performance. The performance of SURF under blur invariance is not sufficient. SURF is also very sensitive to viewpoint change and illumination conditions [65].

D. HISTOGRAM OF ORIENTED GRADIENTS (HoG)

Object detection using HoG descriptors is a hot area [52], which achieved extensive success in the field of human detection. Unlike SURF [51] and SIFT [50], HoG are global descriptors as opposed to local descriptors. The HoG method consists of vector space that calculates similarity employing Euclidean or cosine distances, which is well suited to machine learning techniques. The HoG features construct the gradients inside a cell into nine orientation bins without discriminating between boundary edges and internal textures.

The resulting histogram of oriented gradients (HoG) is used to represent an object category, and constitutes a filter for detecting a particular class of object. The HoG filter is used as a template to detect a particular object in the image, by applying it to various scales of an image pyramid.

The detector can be considered as a classifier that uses an image, a location within that image, and a scale as inputs. The classifier decides if there is an instance or not at the specified location and scale of the target category. As the method is a simple filter, a score can be computed as the scalar product $\beta \cdot \phi(x)$ where β defines the filter, x describes an image with an identified location and scale, and $\phi(x)$ delineates a feature vector – constructed by computing the histogram of gradients from the image x.

A primary innovation of the Dalal-Triggs [52] detector is its ability to extract features that are particularly relevant.

Success: The algorithm is structured with low complexity. It also contains all the gradients of summation in a single cell without taking their magnitude into account [73].

Challenges The algorithm performed well in case of human but could not perform well on detecting objects with high articulation such as birds [74].

1) HoG WITH LBP

Many enhanced, recent techniques use HoG features when detecting human class. One of them is HoG and LBP (Local Binary Pattern) combination [75]. In this method HoG features are extracted following [52]. Besides, pattern histograms are directly built in cells for the construction of the cell-structured LBP. Then the current scanning window and the histograms of LBP patterns from different cells are combined to describe the texture. Two types of detectors are used in this model; one is global detector and the other is part detectors. Global detectors scan the entire windows and the local regions are detected by part detectors. For this training data are provided for learning using linear SVM. For each ambiguous scanning window, this model takes the response of each block of the HoG feature to the global detector to construct an occlusion likelihood map. Then Meanshift approach segments the occlusion likelihood map. The segmented part of the window is determined as an occluded region which responses are mostly negative. In the un-occluded regions, part detectors are applied for finishing in the current scan window if a partial occlusion is seen in a particular scanning window.

a: SUCCESS

The combination of HoG with LBP in the framework of integral image can detect human reasoning with partial occlusion while providing high accuracy [75], [76].

b: CHALLENGES

This model cannot handle the articulated deformation of people [75].

2) HoG WITH CSS FEATURE

Combination of HoG with color self-similarity termed CSS was proposed in [85]. This technique shows high accuracy in one or multiple sequencial images. CSS determines the color patterns in persons which captures pairwise statistics of spatially contained color distributions.

A calculation of the similarity between color histograms of two cell regions is called the Color Self-Similarity (CSS) feature. By calculating the color similarity of two local regions, this feature takes the similarity or dissimilarity with the detection object. Color similarity is used as a feature and gives the benefit that the detection object color is independent. Normalization is involved in computational procedure. The input image is converted to a pre-defined dimension and divided the image into cells, b, of $N \times N$ pixels. After that, color histograms are generated for the cells illustrated in the HSV color method (H: hue, S: saturation, V: brightness) [92]. The color similarity between the two cells uses the color histograms generated from the two cells. Bhattacharyya distance and the intersection of histogram are employed to measure the distance between histograms. Euclidian distance is also adopted to measure the similarity calculations.

a: SUCCESS

HoG with color self-similarity named CSS improves state of-the-art detection performance for both static images and image sequences. CSS determines the color patterns in individuals that capture statistics from spatially contained color distributions in pairs [85].

b: CHALLENGES

This model cannot handle occlusion [85]. The filtering strategy is used which is too stringent and resulted in under reporting of detector performance [93]. In spite of using optical flow, there is still no evidence for the improvement of detection per frame on particular monocular picture frames [39].

TABLE 2. A brief summary of HoG, HoG with LBP, HoG with CSS and Efficient HoG techniques.

methods	application	advantages	limitations
HoG	object detection [52], pedestrian detection [77], text recognition [78]	• invariant to photometric and geometric transformation [52], [79] • improves the detection accuracy and speed [52] • invariant to illumination condition or shadowing [80]	• cannot detect high articulated object perfectly [74] • spatial neighboring pixels context are missed [81]
HoG with LBP	object detection [75], face recognition [82], hand gesture recognition [83], pedestrian detection [84]	• can handle partial occlusion [75], [76] • improves the detection accuracy [75], [76]	• unable to handle the articulated deformation of the object [75]
HoG with CSS	object detection [85], [86], pedestrian detection [87]	• improves the classification accuracy for both static images and videos [85]	• unable to handle occlusion [85]. • no evidence for the improvement of detection per frame on particular monocular picture frames [39].
Efficient HoG	object detection [88], pedestrian detection [88], [89], action recognition [90]	• reduces the computational time [88]	• cannot solve the occlusion issue [88] • unable to perform in real-world environment [91]

3) EFFICIENT HoG HUMAN DETECTION

Another HoG based model has been proposed by [88] with high classification accuracy. This model is capable of computing the HOG features efficiently. At the beginning this algorithm calculates the block-based HOG features at one time. Besides, it reuses the features for all the detection widows to intersect at the block. After dividing each cell into four sub-cells in a block, it classifies and treats them differently into three forms. This transforms cell-based trilinear interpolation to sub-cell-based interpolation. The trilinear interpolation on the basis of sub-cells prevents unimportant interpolation by removing gradients. The look-up-table trick is utilized to accelerate the computation process in the last step.

a: SUCCESS

This model reuses the block-based HoG features for all intersecting detection widows and also utilizes sub-cell based interpolation for reducing the computational time [88].

b: CHALLENGES

This model focuses only on the static images and cannot solve the occlusion problem. Authors capture images in top view only for reasoning occlusion [88]. It is not suitable for identification precision and computational performance in real-world environments [91].

Table 2 illustrates a brief evaluation of HoG, HoG with LBP, HoG with CSS and Efficient HoG methods.

E. BAG OF WORDS (BoW)

Bag-of-Words (BOW) has attracted the research community because of its robustness and simplicity in object recognition.

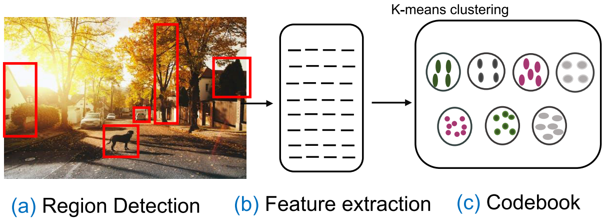

FIGURE 7. Building bag-of-words for image representation; beginning step is to detect the region (a), next step is to extract the features (b), codebook construction is the final step.

The BoW model is generally used in document classification. In 2005 BoW [94] has been used in computer vision because of its robustness and simplicity.

BoW is based on LDA (Latent Dirichlet Allocation) method. A category variable has been introduced in BoW model for image classification. A single image is considered as a group of local patches. Codewords illustrate the patches separately and create a big vocabulary. This is a generative Bayesian hierarchical model for creating an image in a particular category. The process is in general English. Initially, a category label is chosen for an example a mountain image. A probability vector is drawn for the mountain class and it will choose intermediate theme(s) at the time of creating all patches of the image. At first a specific theme is chosen from the mixture of possible themes for making all patches in the scene. Let a ‘‘rock’’ theme is chosen, so the codewords which repeatedly appear in rocks will get advantage. After that if horizontal edges supporting theme is selected, codeword seems to be a horizontal line portion. The procedure is repeated multiple times to create themes and codewords.

Finally, a bag of patches is formed which create an image of mountains. It is also called Theme Model 1. The complete generative calculation of this algorithm can be written by the joint probability.

$$p(x, z, \pi, c|\theta, \eta, \beta) \\ = p(c|\eta)p(\pi|c, \theta).\Pi_{n=1}^{N} p(z_n|\pi)p(x_n|z_n, \beta) \quad (15)$$

where x stands for patches, z indicates for theme, c represents category variable, π is multinomial variable, θ is Dirichlet parameter.

$$p(x|\theta, \beta, c) \\ = \int p(\pi|\theta, c)\left(\Pi_{n=1}^{N} \sum_{z_n} p(z_n|\pi)p(x_n|z_n, \beta)\right)d\pi \quad (16)$$

This is intractable because of the combination into π and β.

BoW is based on some steps: key-points detection, local descriptors, codebook construction (shown in Figure 7).

Key-points detection: Local interest regions or points detection in pictures is the first step of BoW methodology (shown in Figure 7a). For extracting features of key-points, they are calculated at predefined positions and scales. To extract local regions four different algorithms have been examined- Evenly Sampled Grid, Random Sampling, Kadir and Brady Saliency Detector, Lowe's DoG Detector.

Local Descriptors: The second step is to compute local descriptors for the detected keypoints. Pixel grayvalue and SIFT representation have been used to describe a patch. SIFT representation is more robust than the pixel grayvalue representation. The dimensionality of the SIFT descriptor is 128 and Pixel grayvalue is 11 × 11.

Codebook construction: In the next step, the training images of all categories which are the collection of detected patches are learned in the codebook (shown in Figure 7c). The BOW method discards all spatial information about how features are related and distributed across images. BoW features extraction depends on vector quantization. In general, the clustering algorithm k-means is commonly employed for this assignment, and the amount of visual words depend on the clusters(i.e., k) number. The use of BoW can automatically and without guidance learn appropriate intermediate representations of scenes. Texton histogram models can be established and expanded easily using BoW.

Success: BoW can create groups of objects; for instance, humans into a responsive hierarchy [94].

Challenges: The computational cost of BoW model is high [95]. Vector quantization reduces the discriminative power of images and the BoW methodology ignores geometric relationships among visual words [96]. The image interface carries mixed image data that can contain multiple artifacts and background, and such noisy (or diluted) feature representations can reduce the annotation accuracy [97].

1) SPATIAL PYRAMID MATCHING(SPM)

"Beyond Bags of Features" [98] provides a way to identify categories of scenes developed on estimated global geometric correspondence. This method performs by splitting up the picture within each sub-region into progressively fine sub-regions and calculating histograms of regional characteristics.

To detect interest point authors used two sorts of features: weak features and strong features. Weak features are aligned at edge points that are extracted on two scales and eight directions, for a total of $N = 16$ channels. Strong features are used to get better discriminative power. A dense regular grid is used to catch uniform regions instead of interest points. K-means clustering of a random part of patches from the training set is implemented to construct a visual vocabulary. Typical vocabulary sizes for this experiments are $N = 200$ and $N = 400$.

Pyramid matching process has been done by assigning a series of progressively coarser grids over the space of the feature and by taking a weighted sum of the number of matches at each resolution point. When two points are paired at any stable resolution, they fall into one grid cell. The result of matching at finer resolutions is stored lower than matches at higher resolutions.

Order-]less image representation can be handled by pyramid match kernel, which allows precise matching of two collections of features in a high dimensional appearance space, but discards all spatial information. Pyramid matching performs in the two-dimensional image space and applies traditional clustering methods in feature space. Every feature vectors are quantized into N discrete types, and produce the simplifying assumption that only features of the same type can be matched to one another. The final execution concern is normalization. Histograms are normalized altogether by the full weight of image features for maximum computational proficiency and to keep the same number of features in all images in weight forcing. For multi-class classification a support vector machine (SVM) is trained with the help of one-versus-all rule: a classifier is taught to distinguish each class from the other and the classifier label with the highest response is assigned to the test image.

a: SUCCESS

SPM is an easy and computationally effective addition of bag-of-features scene representation which shows high performance on challenging image categorization assignments [98].

b: CHALLENGES

One of the challenges is that the spatial pyramid matching (SPM) method' weights mechanisms is not sophisticated enough. SPM does not work well for the coarse resolution blocks by assigning less weight. In addition, for the finer resolution block containing only background or clutter, it misleads the calculation by assigning more weight [99]. SPM often fails to offer sufficient discriminative power, as seen from the similar image statistics [100]. SPM is built on top pooling which is the most popular spatial pooling method for classifying objects. This gives the irrational features result of the picture when the object of interest (here a car) comes out in pictures at different positions. For this reason, it becomes more challenging to train an appearance method of the object [101].

2) BAG OF VISUAL WORDS (BOVW)

Bag of visual words (BOVW) [108] has the same concept of bag of words (BOW). However, in BOVW image features are used as the ''words'' instead of words. Image features can be detected in an image that are distinctive in shape. To describe an image as a set of features is the key concept of visual words (BOVW). Features consist of keypoints and descriptors. Keypoints are salient points in an image and remain unchanged under the condition of image rotation, shrink or extension. Descriptors detect the keypoints. Difference of Gaussian (DoG) detector is used to detect keypoints automatically in images. The identified keypoints are depicted employing PCA-SIFT descriptor, that is a 36-dimensional real-valued feature vector. Vector quantization (VQ) technique is used to cluster the keypoint from the descriptors. For clustering purpose K-means algorithm is used. Every single cluster is considered as a visual word that illustrates a particular local pattern shared by the keypoints in that cluster. Therefore, the clustering procedure creates a visual-word vocabulary defining various local patterns in images. Finally, for each image, frequency histogram from the vocabularies and the frequency of the vocabularies in the images are created. Those histograms are bag of visual words (BOVW). Image category prediction or corresponding images can be found from the frequency histogram. The representation of bag-of-visual-words can be changed into a visual-word vector similar to the term vector of a document. Visual-word vectors are applied in image classification method.

a: SUCCESS

Bag of visual-words representation produces good classification performance [108].

b: CHALLENGES

The performance is not relatively better as compared to the best standard approaches such as color histograms and homogeneous texture. The spatial location of visual words is not included in the image while using the BOVW approach. This decline in output provides the evidence that the visual terms are too discriminatory for very large dictionaries, i.e. they no longer avoid image disruptions from noise, blurring and discretization [109]. The number of occurrences of each visual word is limited and the computational cost of the visual vocabulary is high [111].

Table 3 illustrates a brief evaluation of BoW, SPM and BOVW methods.

F. DEFORMABLE PART MODEL (DPM)

One of the most successful approaches in general object detection Deformable Part Model (DPM) is basically built on the pictorial structures framework. DPM [112] employs a scanning window technique that consists of a comprehensive ''root filter'' shown in (Figure 8a) and multiple ''part filters'' shown in (Figure 8b). Part model defines a spatial model individually and part filter as well as the spatial model establish a group of permissible placements for a part corresponding to a detection window and the amount of deformation for individual position. The detection window result is similar to classical part-based methods, ''root filter'' as well as ''part filters'' are recorded by calculating the dot product within a window between a set of weights and gradient histogram (HoG) features. The ''part filters'' compute features two times of the spatial resolution of the ''root filter''. Here ''root filter'' behaves similar to a Dalal-Triggs method. The outcome of this method at a specific point and scale in a picture is the result of the ''root filter'' on the window plus the sum over parts, of the limit over positions of that component, of the ''part filter'' record on the coming from sub-window minus the cost of deformation. The task of DPM is to detect objects by looking at a picture pyramid; it is defined at a fixed scale.

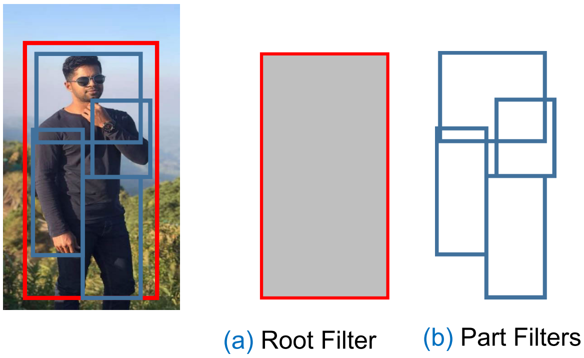

FIGURE 8. Human detection with Deformable Part Model, (a) shows the root filter and (b) shows the part filters.

For an object a model has a root filter M_0 and n part models (M_i, v_i, d_i) where M_i defines ith ''part filter'', v_i stands for two dimensional (2D) anchor position of ith part relative and d_i describes four dimensional deformation parameter for ith part. The position of every filter in the model in a feature pyramid is specified by an object hypothesis, $z = (q_o, \ldots, , q_n)$ where $q_i = (x_i, y_i, l_i)$ defines the level and location of the ith filter. Hypothesis score depends on the score of every filter at its particular position minus a deformation cost that depends on the function of displacement of the part from anchor position plus standard bias,

$$score(q_o, \ldots, q_n) = \sum_{i=0}^{n} M_i' \cdot \sigma(H, q_i) - \sum_{i=1}^{n} d_i \cdot \sigma_d(dx_i, dy_i) + b, \quad (17)$$

where

$$(dx_i, dy_i) = (x_i, y_i) - (2(x_o, y_o) + v_i) \quad (18)$$

TABLE 3. **A brief summary of BoW, SPM and BOVW techniques.**

methods	application	advantages	limitations
BoW	object detection [94], [102], object recognition [103], text classification [104], image retrieval [105]	• quite simple to comprehend and implement [94] • can categorize the objects [94]	• computationally expensive [95] • skips geometric relationships among visual words [96] • less annotation accuracy [97]
SPM	object recognition [106], image classification [98], [107]	• computationally effective [98] • improves the classification accuracy [98]	• weight's mechanism is not sophisticated [99] • insufficient discriminative power [100]
BOVW	scene classification [108], [108], land-use classification [109], object classification [110]	• improves the classification accuracy [108]	• visual vocabulary's computational cost is high [111]

gives the displacement of the ith part relative to its anchor position and

$$\sigma_d(dx_i, dy_i) = (dx, dy, dx^2, dy^2) \tag{19}$$

are deformation features. The score of a hypothesis z can be described in respect of a scalar product, $\delta \cdot \theta(H, z)$, between a vector of model parameters δ and a vector $\theta(H, z)$,

$$\delta = (M'_o, \ldots, M'_n, d_1, \ldots, d_n, b), \tag{20}$$

$$\theta(H, z) = (\sigma(H, q_o), \ldots \sigma(H, q_n), -\sigma_d(dx_1, dy_1), \ldots, -\sigma_d(dx_n, dy_n), 1). \tag{21}$$

This represents a relation within models and linear classifiers. In training, a set of annotated images is provided with bounding boxes on all sides of every instance of the object.

DPM brings the detection issue to a binary classification problem. Every instance x is recorded by the form function,

$$m_\delta(x) = \max_{z \in Z(x)} \delta \cdot \Sigma(x, z). \tag{22}$$

where δ defines a vector of model parameters and z defines latent values (e.g. the part placements). A latent variable MI-SVM construction called latent SVM (LSVM) is used to train models using partly labeled data. A significant property of LSVMs is that if the latent values are set for positive instances, the training problem converts into convex. This is used in an algorithm of coordinate descent.

Traditional SVM training is applied in practice to triples $(x_1, z_1, y_1``, \ldots, x_n, z_n, y''_n)$ where z_i defines the highest recording latent label of x_i in the earlier iteration model. Bounding boxes of PASCAL data-set generates an initial root filter and the parts are initialized from this root filter.

DPM provides an elegant framework for object detection and recognition. DPM can handle Non-rigid deformations for example it can detect human in different poses wearing different clothes. DPM can deal with intra-class variation in features and other graphical properties for instance it can detect cars in different forms and colors. *Success:* It is a better representation of objects. In practice, DPM maintains a list of vectors, $\cdot\Sigma(x, z)$, in lieu of (x, z) pairs. This illustration is easy (it is application free) and more impenetrable [112].

Challenges: Deformable part-based methods obtain state-of-the-art achievement for object identification; yet it is not salient because it depends on heuristic initialization during training for the sake of the optimization of non-convex cost function. Non-convex optimization is unstable to initialization [113]. DPM and its variants are systematically outmatched by methods using a single component and no parts casting doubt on the need for parts. For walking human detection there is still no clear evidence for the necessity of components and parts, beyond the case of occlusion handling [39].

1) SHARED PARTS METHOD

Shared parts method [114] is an addition of DPM that offers for allocation of object share models among several mixture components as well as object classes. In this shared parts method 1) DPM is reformulated to organize part allocation, and 2) a new energy function is proposed for the combined training of mixture parts and object categories. In this extension of DPM model, parts are shared among mixture components. The mixture parts match with various viewpoints in an object. Parts are appeared in multiple mixture components which are visible from a range of views. This sharing parts representation allows the learning of robust models specified a limited training set. Moreover, sharing parts enable stronger methods to be learned in the sense that it becomes achievable to precisely model 'intermediate' graphic modes, for example three-quarters observations of an object. 'Combination weights' β: are acquainted for each component of the mixture. The part filters are shared while spatial priors and anchors are not shared. As a result, spatial priors and anchors are connected with each other to a certain mixture component. Learning a model involves guesstimating $W, V, \ \beta$ parameters so that they simplify well to unseen data; where W describes the appearance and V describes the spatial configuration of the parts. Contrariwise, β describes linear combination of parts in the mixture components. An energy function E is expressed to learn the mixture components which consists of two parts: a regularization term R and a loss term. Regularization term R ensures that the acquired detector's proper observation of unseen data and loss term calculates the training data's detector prediction ability. Energy function offers for concurrent learning of multiple mixtures as $h(\cdot)$ constantly select the component of the mixture that meets the inference scheme. By redeveloping the loss term, the energy $E(\cdot)$ can be extended to multi-class learning for collecting the losses of multiple detectors. Such detectors exchange partial responses based on the weights they have learned from the combination. The regularization term $R(\cdot)$ controls over-fitting by soft-bounding the mixture component responses. All values of β are forced to be positive and weights w are combined. Therefore, regularization needs the penalization of the combined entity βw. For multi-class learning, the regularization definition is simply expanded by summation across whole object detectors.

a: SUCCESS

The total parameter numbers are decreased and training samples are distributed across all applicable parameters, resulting in less negative impact on the lack of available training data. In this method computational expenditure on training and testing time is decreased and enables a large number of classes to be scaled [114].

b: CHALLENGES

Object part models are shared among multiple mixture components and object classes and this introduces additional noise [115]. In Fine-Grained Video Classification (FGVC) similarities between classes are exploitable for model sharing methods [116].

2) STRONGLY-SUPERVISED DPM

This model [113] extends DPM methods using additional supervision. It is necessary to know linear parameters of the model β for this model training samples and annotations D and C number of clusters. Positive samples are clustered based on their pose, every mixture component is assigned to a structure and part filters are initialized. Mixture methods (components) allow intra-class modification modeling in the presence of various examples. These changes are caused by sturdy variations in viewpoints, class subcategories and unstable deformations. Assigning positive examples to LSVM elements is a non-convex expansion and therefore unstable to initialization. For better assignment of training samples part annotations is used to have components according to the pose. As a result, similar parts are aligned better inside every component while employing linear classifiers. Annotated parts are used to parametrize the pose θ_x of sample x. All positive pose vectors are clustered by applying a modified k-means clustering algorithm. A weight is defined for all blocks of θ by a predetermined parameter (W) to monitor the consequence of all parameters for the time of clustering. The position of a bounding box p_i in the picture and the status of the binary visibility v_i are described for all parts of this method. One mixture component of the method has a tree structure with grids L and edges M related to object parts and relations among parts accordingly. An optimization process is designed for making a dependency graph employing part annotations of object. The statistically optimal component relationship of the generative pictorial structure method can be accomplished by optimizing the connections over the prior probability of samples.

Then a fully connected graph $H = (L_H, M_H)$ is constructed with $n + 1$ grids with respect to object and parts bounding boxes, each edge m is described as a 3-tuple (i, j, w). For every pair of grids (i, j), the translation's diagonal covariance matrix (dx, dy) is calculated between the two related bounding box centers. Optimization problem is solved by following a coordinate-descent method in which at the start the latent variables for positive records are set in order that the equation turns into convex in β. The stochastic gradient descent (SGD) is then used to estimate parameters of the model. Object portions are often occluded because of the existence of additional objects and self-occlusions. As occlusions mostly do not occur arbitrarily, there may have a compatible presentation of occluded parts positions. For occluded parts occlusions model is constructed by learning separate appearance parameters F^o. The bias terms b_i and b_o^i manage the stability in S_A between occluded and non-occluded presentation. An improved Latent SVM is used to train models by collecting hard negative records from negative training pictures.

a: SUCCESS

A learning design is formulated that can cope with annotations of sub-optimal and incomplete regions of objects. Besides, it deals with partial occlusions of objects explicitly [113].

b: CHALLENGES

The occluded part in supervised DPM is modeled on additional templates and are based upon a root portion (i.e. the holistic object), which never becomes "occluded." If a major occlusion occurs, it is difficult to model the root component itself [117]. The part models mentioned provide prototypes for an occluding version of each component of the model. However, the visibility of different parts of the model is not enforced by pairs [118].

3) WEIGHTED HISTOGRAMS OF ORIENTED GRADIENTS(WHoG)

Weighted Histograms of Oriented Gradients (WHoG) [73] is another compatible descriptor to detect object. Weighted Histograms of Oriented Gradients (WHoG) is a join of global shape descriptors and local point descriptors that descript object detection with diverse textures. Authors use bottom-up pose clustering to handle intense pose variations. One is input pose taker templates and another is taker of color information;

TABLE 4. A brief summary of DPM, Supervised DPM and WHoG techniques.

methods	application	advantages	limitations
DPM	object detection [112], [119], pedestrian detection [120], face detection [121], pose estimation [122], image segmentation [123]	• better object representation and application free descriptor [112] • can handle partial occlusions [112] • improves the accuracy for object detection (achieves the state-of-the-art) [112], [119]	• computationally expensive [124] • unstable to initialize the non-convex optimization [113]
Strongly-Supervised DPM	object detection [113], object classification [113]	• can handle partial occlusion [113] • improves the detection accuracy [113]	• problematic to model the root component for major occlusion [117]
WHoG	object detection [73]	• improves the detection accuracy [73] • can detect textured objects in opposition to background clutter [73] • minimizes computational complication [125]	• unable to meet the requirement for real-time processing [126]

two CNNs works parallel there. Firstly, the authors explain pose clustering technique and employ an improved HoG descriptor (WHoG) to make a pose-specific bird detector. WHoG generates a result for every candidate posture, with the maximum score indicating the recognized position for that specific posture. In the subsequent step, the authors use scale invariant color components for making a spices-specific bird detector. WHoG finds out the birds and defines its position exactly in any particular picture by allocating additional weight to fringe and less load to body structures or stripes. WHoG can be usable to detect any kind of object that conveys intestinal designs or textures.

a: SUCCESS

This approach can detect textured objects in the presence of background clutter. Moreover, it has an immense dimensional articulation as well as pose variety, for instance, in birds. Integration of properly schematic scale invariant color features into the method increases the detection accuracy. This method minimizes computational complication and shows a great performance progress on a comprehensive data-set: CUB-200-2011 [125].

b: CHALLENGES

It is difficult to integrate WHoG straight into FPGA for quick object detection as it is not completely adapted for resource-limited hardware systems [127]. WHoG is designed for a CPU that does not consider the requirement for real-time processing [126].

Table 4 illustrates a brief evaluation of DPM, Supervised DPM and WHoG methods.

G. ORTHOGONAL MOMENTS

Orthogonal moments have a key role to play in image processing and other related applications. Over the last few decades, orthogonal moments have been extensively applied in different fields in image analysis [128]. Moment invariants and moments have the potential to describe global image processing features.

Moments can be defined as the projections of an image function onto particular kernel functions [129], in Rectangular coordinates:

$$Z_{nm}^{c} = \int_{R^2} h_{nm}(x, y) f(x, y)\, dxdy \tag{23}$$

or in polar coordinates:

$$Z_{nm}^{p} = \int_0^1 \int_0^{2\phi} h_{nm}(r, \theta) f(r, \theta)\, rdxrd\theta \tag{24}$$

where Z_{nm}^{c} indicates the $(n+m)th$-order moments of an image function $f(x, y)$ represented in Rectangular coordinates, Z_{nm}^{p} stands for the $(n+m)th$-order moments of an image function $f(r, \theta)$ represented in polar coordinates $n, m = 0, 1, 2, \ldots,$ and $h_{nm}(x, y)$ and $h_{nm}(r, \theta)$ are the kernel functions (also known as the basis functions). With different kernel functions, different types of moments can be obtained.

The moments are orthogonal moments if those kernel functions satisfy the following conditions

$$\int_0^1 \int_0^1 h_{nm}(x, y) h_{st}(x, y)\, dxdy = k_1 \delta_{ns} \delta_{mt},$$
$$\int_0^1 \int_0^{2\phi} h_{nm}(r, \theta) h_{st}(r, \theta)\, rdrd\theta = k_2 \delta_{ns} \delta_{mt} \tag{25}$$

where k_1 and k_2 are the normalization coefficients and δ_{ns} is the Kronecker delta function; otherwise, the moments are called non-orthogonal moments.

Success: With their lower order (not larger than 12th order) elements, orthogonal moments are strong signal descriptors that have distinguishing strength in object or pattern recognition [129].

Challenges: The major drawback of orthogonal moments is the difficulty of their computation [129].

1) KRAWTCHOUK POLYNOMIALS

Krawtchouk polynomials are utilized for the formation of discrete moments that can illustrate a picture locally [23]. Discrete Krawtchouk polynomials are commonly used in various areas for their exceptional localization property and characteristics [130].

The definition of the n-th order classical Krawtchouk polynomial [130] is defined as

$$K_n(x;p,N) = \sum_{k=0}^{N} a_{k,n,p}x^k = {}_2F_1(-n,-x;-N;\frac{1}{p}) \tag{26}$$

where $x, n = 0, 1, 2, \ldots, N$, $N > 0$, $p\epsilon(0, 1).{}_2F_1$ is the hypergeometric function, defined as

$${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \frac{(a)_k(b)_k}{(c)_k}\frac{z^k}{k!} \tag{27}$$

and $(a)_k$ is the Pochhammer symbol defined by

$$(a)_k = a(a+1)\ldots(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)}. \tag{28}$$

The set of $(N + 1)$ Krawtchouk polynomials $K_n(x; p, N)$ forms a complete set of discrete basis functions with weight function

$$w(x;p,N) = \binom{N}{x} p^x(1-p)^{N-x} \tag{29}$$

and fulfills the orthogonality condition

$$\sum_{x=0}^{N} w(x;p,N)K_n(x;p,N)K_m(x;p,N) = \rho(n;p,N)\delta_{nm} \tag{30}$$

where $n, m = 1, 2, \ldots, N$ and

$$\rho(n;p,N) = (-1)^n\left(\frac{1-p}{p}\right)^n \frac{n!}{(-N)_n}.$$

a: SUCCESS

Krawtchouk polynomials consistently perform better compared to other polynomials for reconstruction error [23], [130].

b: CHALLENGES

A particular window function is required for signal. The computational time of krawtchouk polynomials is relatively high [23].

2) TCHEBICHEF POLYNOMIALS

The Tchebichef polynomials have been successfully applied as pattern characteristics in two-dimensional (2D) image analysis [131], [132].

The discrete Tchebichef polynomials [131] are defined as

$$t_n(x) = (1-N)_n \;\; {}_3F_2(-n,-x,1+n;1,1-N;1), \qquad n,x,y = 0,1,2,\ldots N-1 \tag{31}$$

where $(a)_k$ is the Pochhammer symbol expressed by

$$(a)_k = a(a+1)(a+2)\ldots(a+k-1) \tag{32}$$

and ${}_3F_2(\cdot)$ is the generalized hypergeometric function

$${}_3F_2(a_1,a_2,a_3;b_1,b_2;z) = \sum_{k=0}^{\infty} \frac{(a_1)_k(a_2)_k(a_3)_k}{(b_1)_k(b_2)_k}\frac{z^k}{k!}. \tag{33}$$

The Tchebichef polynomials fulfill the property of orthogonality with

$$\begin{aligned}\rho(n,N) &= \frac{N(N^2-1)(N^2-2^2)\ldots(N^2-n^2)}{2n+1}\\ &= (2n)!\begin{bmatrix} N+n \\ 2n+1 \end{bmatrix}, n = 0,1,\ldots,N-1\end{aligned} \tag{34}$$

and have the following recurrence relation:

$$\begin{aligned}&(n+1)t_{n+1}(x) - (2n+1)(2x-N+1)t_n(x)\\ &\quad + n(N^2-n^2)t_{n-1}(x) = 0, \quad dn = 1, 2\ldots, N-1.\end{aligned} \tag{35}$$

a: SUCCESS

This feature perfectly eliminates the necessity for any numerical approximation and satisfies the orthogonal property precisely in the image discrete domain [131], [133].

b: CHALLENGES

The TP's common issue is that the calculation of coefficients is vulnerable to numerical instability for higher polynomial order [132].

3) CHARLIER POLYNOMIALS

Charlier polynomials (CHPs) [134] are generally adapted in image analysis [135] because of their excellent success in the study of signal processing and their potential for signal representation [136].

The definition of CHP $C_n^p(x)$ for the n-th order is obtained from [134], [136] as follows:

$$C_n^p(x) = {}_2F_0\left(\begin{matrix}-n,-x\\ -\end{matrix} \;\middle|\; -\frac{1}{p}\right) \qquad n,x = 0,1,\ldots.N; p \geq 0 \tag{36}$$

where p indicates the parameter of the CHP, and ${}_2F_0$ is the hypergeometric series which is expressed by [136], [137]:

$${}_2F_0\left(\begin{matrix}a,b\\ -\end{matrix} \;\middle|\; -z\right) = \sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{k!}(z)^k \tag{37}$$

where $(a)_b$ denotes the ascending factorial, which is also known as the Pochhammer symbol, and expressed as follows:

$$\begin{aligned}(a)_b &= a(a+1)(a+2)\ldots(a+b-1)\\ &= \frac{(a+b-1)!}{(b-1)!} = \frac{\Gamma(a+b)}{\Gamma(b)}\end{aligned} \tag{38}$$

On the basis of Equations (36) and (37), CHP can be defined as:

$$C_n^p(x) = \sum_{k=0}^{\infty} \frac{(-n)_k(-x)_k}{k!}(-\frac{1}{p})^k \tag{39}$$

CHP satisfies the orthogonality conditions such that [136], [138]:

$$\sum_{x=0}^{N} C_n^p(x)C_m^p(x)w_c(x;p) = \rho_c(n;p)\delta_{nm} \tag{40}$$

where $w_c(x;p)$ and $\rho_c(n;p)$ are the weight function and squared norm of CHP $C_n^p(x)$, respectively. The weight function is expressed as follows [136], [139]:

$$w_c(x;p) = \frac{(e)^{-}p(p)^x}{x!} \tag{41}$$

and the squared norm $\rho_c(n;p)$ is defined as follows [139]:

$$\rho_c(n;p) = \frac{x!}{(p)^n} \tag{42}$$

a: SUCCESS

This approach minimizes both the time of computation and the error of propagation [136]. The properties of this model can be utilized to compact a natural image and to recreate a large image [134].

b: CHALLENGES

The coefficient's numerical inconsistency for higher-order polynomials is the key problem of CHPs [136].

4) SQUARED KRAWTCHOUK–TCHEBICHEF POLYNOMIAL (SKTP)

Two orthogonal polynomials' combination is also considered as orthogonal polynomial. Motivated by this idea Abdulhussain *et al.* presented one new hybrid technique for object detection named SKTP (squared Krawtchouk–Tchebichef polynomial) [140].

From the mathematical point of view, a polynomial established by combining two OPs is also orthogonal [141], [142]. The nth order of the hybrid polynomial form, $R_n(x)$, can be defined as follows:

$$R_n(x;N) = \sum_{j=0}^{N-1} X_j(x;p,N)Y_j(n;p,N)$$
$$n, x = 0, 1, \ldots, N-1 \tag{43}$$

where $Y_n(x;p,N)$ and $X_n(x;p,N)$ are orthogonal polynomials constructed from two fundamental OPs, i.e., first level of combination. Let $X_n(x;p,N)$ and $Y_n(x;p,N)$ be expressed as follows:

$$X_n(x;p,N) = \sum_{i=0}^{N-1} K_i(n;p)T_j(x) \tag{44}$$

$$Y_n(x;p,N) = \sum_{i=0}^{N-1} K_i(x;p)T_j(n)$$
$$n, x = 0, 1, \ldots, N-1; p \in (0,1) \tag{45}$$

From Equations (43), (44), and (45), the developed hybrid OP can be defined as follows:

$$R_n(x;p,N) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{l=0}^{N-1} K_j(i;p)T_j(x)K_l(n;p)T_l(i) \tag{46}$$

a: SUCCESS

The performance of SKTP is stable and remarkable compared with other existing orthogonal polynomials in noisy environments [140].

b: CHALLENGES

There is no clear information provided for handling major occlusion and the articulated deformation of objects [140].

Table 5 illustrates a brief evaluation of Orthogonal moments, Krawtchouk polynomials, Tchebichef polynomials, Charlier polynomials and SKTP methods.

III. CLASSIFIERS

A. NEAREST NEIGHBOUR CLASSIFIER

The K-nearest neighbor(KNN) is one of the simplest and most traditional nonparametric algorithms for classification proposed by [157]. KNN generates a vector of features and outcomes a vector of values. This vector of values goes into a comparator that compares the features vector; feature vector coming from a library of possibilities. By finding the closest match the comparator determines what the object is [158]. KNN classifies a subset of k training samples from a set of stored samples which are the nearest to it. It uses Distance Metric to compute the smallest distance and majority voting rule for the nearest object selection [159].

Let take a training set with N training samples $(X, M) = (x_1, m_1), .., (x_N, m_N)|x_i \in R^d, m_i \in 1, 2, .., P, i = 1, .., N$ where x_i defines the feature vector of ith sample and m_i defines class label and P stands for predefined classes. From a test sample $x' \in R^d$ of training samples X, we can find its k-nearest neighbors $X' = x'_1, x'_2, \ldots, x'_k$ with their resembling labels symbolized as $M' = m'_1, m'_2, \ldots, m'_k$. Euclidean distance $||x - m||$ is used to calculate the smallest distance for finding neighbors. Then majority voting rule is applied to classify the test sample x'.

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^{k} I_{\{M_{(i)}(x)=1\}} \leq I_{\{M_{(i)}(x)=0\}} \\ 1 & \text{otherwise.} \end{cases} \tag{47}$$

$g_n(x)$ is a majority vote among the labels of the k nearest neighbors.

TABLE 5. A brief summary of Orthogonal moments, Krawtchouk polynomials, Tchebichef polynomials, Charlier polynomials and SKTP techniques.

methods	application	advantages	limitations
Orthogonal moments	face recognition [143], object classification [144], [145], object recognition [128], [146], texture retrieval [147]	• strong signal descriptors with low order elements [129]	• computationally expensive [129]
Krawtchouk polynomials	object recognition [130], edge detection [148], object classification [149], image recognition [150]	• better performance for reconstruction error [23], [130]	• high computational time [23]
Tchebichef polynomials	image analysis [131], face Recognition [151], edge detection [132], image retrieval [152]	• eliminates the necessity for numerical approximation in the image discrete domain [131], [133]	• vulnerable coefficients' calculation to numerical instability for higher polynomial order [132]
Charlier polynomials	object recognition [153], image classification [154], image reconstruction [155], object recognition [156]	• minimizes both the time of computation and the error of propagation [136]	• coefficient's numerical inconsistency for higher-order polynomials [136]
SKTP	face detection [140]	• stable in noisy environments [140]	• no clear information for handling major occlusion [140]

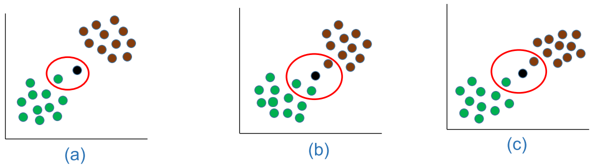

FIGURE 9. K-nearest neighbor to identify objects in class A and B: christi color represents class A, saddle brown represents class B.

Finding of k values is a challenging task. A very large value of K is computationally expensive and prediction accuracy is low for a very small value of K [160]. Besides, KNN is also responsive to lebel noise [161]. We classify a new object from known two classes; A and B (shown in Figure 9). If $k = 1$ (shown in Figure 9a), it will select the nearest one. If $k = 3$ (shown in Figure 9b), it will select the nearest 3 objects. Let the value of K is odd to avoid the voting ties. If $k = 2$ (shown in Figure 9c), (even number) it is difficult to classify the object because of achieving the same score of two classes labels.

B. DECISION TREE

Decision trees are developed for classification tasks. It generates a tree for a data-set and processes a single outcome at every leaf. These trees construction involves a root node at the top then continues down to its leaf nodes.

Iterative Dichotomiser 3 (ID3) is one of them developed by [162]. The basic structure of ID3 is iterative. The sample objects are defined distinctly by a set of properties values where a classification statute is established and decision trees are disclosed successively in respect of these same properties.

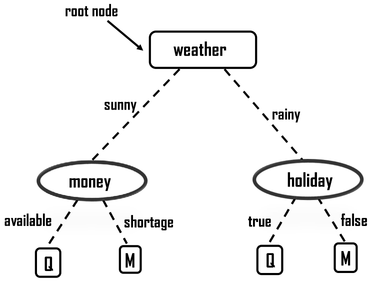

FIGURE 10. Decision tree: starts with the root node and ends with the leaves.

In order to classify an object, ID3 starts at the root of the tree shown in Figure 10, assesses the test, and selects the accurate branch to obtain the result. The process goes on until a leaf is found, at which time the object is assigned to the class named by the leaf.

It means this algorithm performs Top-Down Induction of Decision Trees. The induction creates a decision tree for classifying the objects properly from the training set. Besides, it can classify the unseen objects. There should be some significant relationship between the class of an object and its attribute values for object classification. The best attribute of the data-set is placed at the root of the tree. It splits the training set into subsets. A subset of a training set known as a window is selected randomly and a decision tree is constructed from it; this tree classifies every object in the window accurately. Subsets should be made in such a way that each subset contains data with the same value for an attribute.

Next, in the training data-set each of other objects are categorized employing the tree. If the tree provides the accurate response for each of these objects, it is perfect for the whole training data-set and then the process ends. If it does not provide an accurate answer, a choice of the inaccurately categorized objects will be compiled to the list and the process will begin. In this manner, after just a certain number of iterations accurate decision trees have been identified for training data-sets.

It is difficult to make a decision tree for a random set objects γ. γ being void or containing just one class object, decision tree becomes the simplest while having a labeled leaf only with that class.

On the other hand, γ contains two or more subsets γ_i which are nonempty and their values are smaller than γ, it splits the subsets to fulfill the one class requirement for a leaf. By this process a decision tree is formed to classify every object properly in γ.

The test selection is the key in order to make the decision tree smooth.

For this reason, root node is chosen by selecting an attribute while a test is restricted for ramification on the attribute values.

ID3 appoints an information based method for attribute selection. Information gain is a popular way to select an attribute. Let $I(q, m)$ be the information in q positive decisions and m negative decisions.

$$I(q, m) = -\frac{q}{q+m} log_2 \frac{q}{q+m} - \frac{m}{q+m} log_2 \frac{m}{q+m} \quad (48)$$

If attribute L with values $[L_1, L_2 \ldots . L_v]$ is employed for the root of the decision tree, it will divide γ into $[\gamma_1, \gamma_a \ldots . \gamma_v]$ where γ_i consists of those objects in γ that have value L_i of L. Let γ_i take on q_i objects of class Q and m_i of class M. $I(q_i, m_i)$ is the required information of γ_i for the sub-tree.

The proposed information needed for the tree with L as root is then gained as the weighted average

$$E(L) = \sum_{i=1}^{v} \frac{q_i + m_i}{q+m} I(q_i, m_i) \quad (49)$$

where the weight for the ith branch is the proportion of the objects in γ that belong to γ_i. The information acquired by branching on L is therefore

$$gain(L) = I(q, m) - E(L) \quad (50)$$

A suitable thumb rule would appear to be selecting the branch attribute on which the most information is obtained. ID3 tests every candidate attributes and selects L to optimize $gain(L)$ and then uses the same process recursively to form decision trees for the residual subsets $\gamma_1, \gamma_2 \ldots . \gamma_v$.

To illustrate this idea, consider the data-set γ in Table 6 and the objects "hangout with friends". Of the 10 objects, 5 are of class Q and 5 are of class M, so the information needed for classification is

$$I(q, m) = -\frac{5}{10} log_2 \frac{5}{10} - \frac{5}{10} log_2 \frac{5}{10} = 1 \; bits \quad (51)$$

TABLE 6. Data-set γ.

Day	Weather	Holiday	Money	Decision
1	sunny	true	available	Q
2	sunny	false	available	Q
3	sunny	true	available	Q
4	sunny	true	shortage	Q
5	rainy	false	available	Q
6	rainy	true	shortage	M
7	rainy	false	shortage	M
8	sunny	false	available	M
9	sunny	true	available	M
10	rainy	false	available	M

Now consider the weather attribute with values [sunny, rainy]. The 6 value of sunny are four from class Q and two from class M, so

$q_1 = 4 \;\; m_1 = 2 \;\; I(q_1, m_1) = 0.9182$

and for rainy

$q_2 = 1 \;\; m_2 = 3 \;\; I(q_2, m_2) = 0.8112$

Consequently, the required information after examining this attribute is

$$E(weather) = \frac{6}{10} I(q_1, m_1) + \frac{4}{10} I(q_2, m_2) = .8759 \quad (52)$$

The gain of this attribute is then

$$gain(weather) = I(q, m) - E(weather) = 0.1241 \quad (53)$$

similarly

gain(holiday) = 0.029

gain(money) = 0.0349

Form this analysis, the information gain of weather is high so it will be the root node and the decision tree is shown in Figure 10.

ID3 algorithm is robust to execute and gives high accuracy in classification tasks [162].

The ID3 algorithm depends on the theory of information gain and attempts to reduce the number of comparisons amongst the training records. ID3 algorithm does not give a measurable optimal result. It can stop working in optimums at the local level. Entropy calculates the information gain that illustrates the measure of uncertainty of a record γ [163].

C. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) was developed by [164] for solving the classification problems. Due to its effectiveness and popularity, SVM has been employed in many object detection methods. SVM is used as a linear binary classifier as well as a non-linear classifier that can be explained by a separating hyperplane [165]. From a set of labeled training data, this method produces an optimal hyperplane that separates

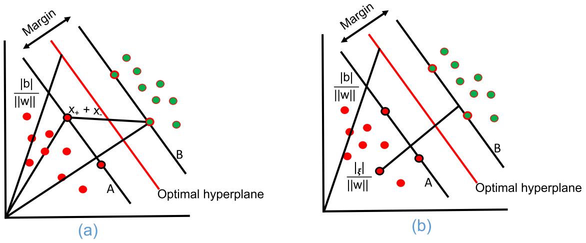

FIGURE 11. **SVM margin construction procedure and decision boundary.**

the positive from the negative examples in two dimensional space.

In Figure 11, some positive examples and some negative examples are shown. Now the question is how to separate these two classes? A straight line called hyperplane or decision boundaries is required to separate them. Another question is how would we make a decision rule that would use that decision boundaries?

Let us take a set of labeled training data $x_i, y_i, i = 1, \ldots.., l$, where $y_i \epsilon -1, 1$, $x_i \epsilon R^d$. To separate the positive from the negative examples some hyperplanes are needed. The equation of hyperplane is

$$u \cdot x + c = 0 \tag{54}$$

where u is normal to the hyperplane, x stands for values lie on the hyperplane, $|c|/||u||$ defines the perpendicular distance from the hyperplane to the origin, and $||u||$ describes the Euclidean norm of u. Some points either positive (x_+) or negative (x_-) are the closest to the hyperplane which are the shortest distance from the separating hyperplane. $(x_+ + x_-)$ is the margin of a separating hyperplane. Support vector algorithm takes the largest margin for the separating hyperplane in linearly separable case. Let us assume that each of the training data satisfies the conditions given by,

$$x_i \cdot u + c \geq +1 \, for \; y_i = +1 \tag{55}$$

$$x_i \cdot u + c \leq -1 \, for \; y_i = -1 \tag{56}$$

The combination of equation (55) and (56) become:

$$y_i(x_i \cdot u + c) - 1 \geq 0 \tag{57}$$

The distance from the nearest positive value or the nearest negative value $= 1 \| |u||$ and their margin is clearly $2 \| |u||$. A and B have no training values in between them which are also described in hyperplane. They are parallel to each other and shown in Figure 11. Based on these conditions the pair of hyperplanes can be found at maximum margin by minimizing $||u||2$, related to constraints (57).

However, the result of two dimensional (2D) cases is described in Figure (11a) in details.

Training values which are closest to the hyperplanes and support the hyperplanes' margin are known as support vectors shown in Figure (11a) with extra circles.

There are some constraints (57); to solve this problem a Lagrangian formulation is employed where positive Lagrange multipliers β_i, $i = 1, \ldots, l$, constraints of the form $d_i \geq 0$; are introduced. To form the Lagrangian, the constraint equations are multiplied by positive Lagrange multipliers and subtracted from the objective function. For equality constraints, the Lagrange multipliers are unconstrained. This gives Lagrangian:

$$L_p \equiv \frac{1}{2}||u||^2 - \sum_{i=1}^{l} \beta_i y_i (x_i \cdot u + c) + \sum_{i=1}^{l} \beta_i \tag{58}$$

This is a convex quadratic programming problem. To solve this L_P should minimize with respect to u, c, where $\beta_i \geq 0$ and "dual" problem also can be solved by maximizing L_P, the gradient of L_P related to u and c vanish give the conditions:

$$u = \sum_i \beta_i y_i x_i \tag{59}$$

$$\sum_i \beta_i y_i = 0. \tag{60}$$

As these are equality constraints in the dual formulation, we can put them into Equation (58) to give

$$L_D = \sum_i \beta_i - \frac{1}{2} \sum_{i,j} \beta_i \beta_j y_i y_j x_i \cdot x_j \tag{61}$$

The optimization depends on the dot product of pairs of samples. In the solution values having $\beta_i > 0$ are defined as "support vectors", and lie on one of the hyperplanes A, B. Support Vectors lie closest to the decision boundary; if all other training values are removed and the training is repeated, the same separating hyperplane would be found. To handle non-separable data positive slack variables η_i, $i = 1, \ldots, l$ are introduced in the constraints, which then become:

$$x_i \cdot u + c \geq +1 - \eta \quad for \; y_i = +1 \tag{62}$$

$$x_i \cdot u + c \leq -1 + \eta \quad for \; y_i = -1 \tag{63}$$

$$\eta \geq 0 \tag{64}$$

where $\sum_i \eta_i$ is defined by an upper bound of the number of having training errors. An additional cost is given to choose the optimal function to minimize from $||u||^2/2$ to $||u||^2/2 + D(\sum_i \eta_i)^k$, where D is the parameter selected by the user, As it stands, this is a convex programming problem for any positive integer k; for $k = 2$ and $k = 1$ it is also a quadratic programming problem, and the selection $k = 1$ has a fruitful effect that neither the η_i, nor their Lagrange multipliers, involve in the Wolfe dual problem, as follows:

$$L_D = \sum_i \beta_i - \frac{1}{2} \sum_{i,j} \beta_i \beta_j y_i y_j x_i \cdot x_j \tag{65}$$

where as:

$$0 \leq \beta_i \leq D, \tag{66}$$

$$\sum_i \beta_i y_i = 0. \tag{67}$$

The given solution is again presented as

$$u = \sum_{i=1}^{Ns} \beta_i y_i x_i \tag{68}$$

The term Ns is known as the number of support vectors. Hence the only main difference from the objective hyperplane consideration is that the β_i have an upper bound of D. This circumstance is described in Figure (11b). To solve the primal issue, Karush-Kuhn-Tucker criteria is required. The primal Lagrangian can be expressed by

$$\begin{aligned} L_p = \frac{1}{2}||u||^2 + D\sum_i \eta_i \\ - \sum_i \beta_i y_i (x_i \cdot u + c) - 1 + \eta_i - \sum_i \mu_i \eta_i \end{aligned} \tag{69}$$

where the μ_i define the Lagrange multipliers which were proposed to effectuate positivity for the η_i. The conditions of KKT for the primal issue are given by

$$\beta_i \{ y_i (x_i \cdot u + c) - 1 + \eta_i \} = 0 \tag{70}$$

$$\mu_i \eta_i = 0 \tag{71}$$

$$0 < \beta_i < D \tag{72}$$

SVM classifies the test sample with the following decision function, which determines on which side of the separation hyperplane the sample x lies,

$$M(x) = sgn(u \cdot x + c) \tag{73}$$

For Nonlinear classification problems, the data are mapped to possibly infinite dimensional Euclidean space H, using a function which is known as π:

$$\pi : R^d \rightarrow H. \tag{74}$$

The training algorithm depends on just the data through scalar multiplications in H, i.e. on maps of the form $\pi(x_i) \cdot \pi(x_j)$. In the training algorithm, we only need to use "kernel function" K where $K(x_i, x_j) = \pi(x_i) \cdot \pi(x_j)$, and would never need to explicitly know what π is. One example is defined as,

$$K(x_i, x_j) = e^{-}||x_i - x_j||^2/2\beta^2. \tag{75}$$

But how can this machine be used? Basically, u is required that will also live in H. However, an SVM is applied in test phase by calculating scalar multiplication of a particular test value x with u, or more generally by evaluating the sign of

$$\begin{aligned} m(x) &= \sum_{i=1}^{Ns} \beta_i y_i \pi(s_i) \cdot \pi(x) + c \\ &= \sum_{i=1}^{Ns} \beta_i y_i K(s_i, x) + c \end{aligned} \tag{76}$$

where s_i define the support vectors. Therefore, $\pi(x)$ calculation can be ignored evidently, rather $K(s_i, x) = \pi(s_i) \cdot \pi(x)$ should be employed.

SVM can achieve optimal result on training record but the VC dimension of SVM is computationally high and hard to calculate [177].

Table 7 illustrates a brief evaluation of KNN, DT and SVM techniques.

IV. APPLICATION

Human Detection is being used in various applications. Human detection constitutes the first phase in a variety of applications for examples "intelligent digital content management" [201]–[203], "driving assistance systems" [204]–[207], "smart video surveillance" [208]–[210], "abnormal behavior" [211], [212], "crowd scene analysis (people counting)" [213], [214], "person re-identification" [215]–[217], "human tacking" [218]–[220], "human activity recognition" [221]–[224], "human pose estimation" [225]–[228], "gender classification" [229]–[232], "pedestrian detection" [233]–[238] and "e-health systems" [239]–[243].

1) ABNORMAL BEHAVIOR

"Abnormal behavior" is defined as any kind of activity that is performed in different or unusual circumstances than what is normally or usually carried out. Abnormal behavior's definition changes according to the circumstances. For instance, a person running in a mall is deemed abnormal but it is considered as normal behaviour in a field. Detection of abnormal behavior is important because of the society's rising crime rate. If an irregular activity can be identified earlier, it is possible to prevent tragedies [244].

Ko *et al.* developed a model for human abnormal behavior detection combined with deep learning using Kalman filter. A LSTM (long-short term memory) model is applied to estimate the behaviour process from normal RGB image [245]. Unlike "trajectory-based" and "pixel-based" methods, Coşar *et al.* came up with one integrated approach while reducing the false alarm for detecting the irregular group behaviors [246]. Benmakrelouf *et al.* combined the "divergence" and "outlier" detection methods to detect the abnormal activities in real time and succeeded with high accuracy rate [247].

2) CROWD SCENE ANALYSIS

"Crowd analysis", that is, people counting in public gathering has become a point of discussion in computer vision [248], [249]. It is considered as one of the major or critical problems in this area. Crowd analysis has shown its importance in some events, for example, marathon terrorist attacks [250].

As "head" is able to be seen clearly in a crowd scene, Shami *et al.* detected the head to analyze the crowd. At first the authors combined the SURF feature with binary SVM classifier to separate the crowd scene from the not-crowd scene. Then the authors detected the heads from these scenes using CNN [251]. Idrees *et al.* came up with one technique to detect humans in deep crowd gathering by employing "Locally-Consistent Scale Prior". Subsequently, the authors

TABLE 7. A brief summary of KNN, DT and SVM techniques.

methods	advantages	limitations
KNN	• has no need for training period that makes the technique to execute faster than other classifiers [166] • possible to seamlessly add new data that would not change the algorithm's accuracy [167] • very simple to implement, only requires two parameters (K value and the function of distance) to implement [157] • less computational time [159]	• difficult to find the K value [160] • less accuracy and high computational time for large dataset [160] • poor performance for high dimensional data [168] • sensitive to noisy environment [161]
DT	• needs less effort in the time of pre-processing for data formulation as opposed to other classifiers [169] • needs no data scaling and normalization [170] • robust to execute and gives high accuracy in classification tasks [162] • not affected by missing values in the dataset while making a decision tree [171]	• minor change in the dataset creates uncertainty in the decision tree structure [172] • computationally expensive [163] • inadequate for calculating continuous values and applying regression [173]
SVM	• provides good classification accuracy [174] • shows good performance for high dimensional data [175] • comparatively memory efficient [176]	• computationally expensive [177] • responsive to noise environment [178] • inferior performance for big dataset [179]

introduced an "Integer Programming formulation" for reasoning occlusion that can enhance the detection localization [26].

3) PERSON RE-IDENTIFICATION

We need to clarify the definition of "re-identification" first to know the "Person re-identification". A. Plantinga defined "re-identification" as "To re-identify a particular, then, is to identify it as (numerically) the same particular as one encountered on a previous occasion" [217]. "Person re-identification (re-ID)" has gained a lot of attention among researchers because of its significance to application and research in the field of computer vision [25]. It attempts to locate a human of interest in another camera [217].

Su *et al.* developed "Semi-supervised Deep Attribute Learning(SSDAL)" technique for human "Re-Identification (ReID)" solving visual appearance changing with high accuracy [252]. Since parts of the human body are often misaligned in the detected person boxes, an image characterization is needed which can deal with this misalignment. Suh *et al.* proposed "Part-Aligned Bilinear Representations" technique to solve this misalignment issue while the authors achieved high accuracy in five different data-sets [253].

4) HUMAN TRACKING

Human tracking is one of the core problems in the field of computer vision. Human tracking is the process in which moving humans are placed in various frames of a video, maintaining the accurate identities. Xu *et al.* developed "UFFB-based INS/UWB-integrated human tracking" technique in which they were able to increase the accuracy of localization [254]. In order to track multiple human, Fu *et al.* proposed "Monte Carlo probability hypothesis density (PHD) filter-based" system [255].

5) PEDESTRIAN DETECTION

Pedestrian detection is another popular topic in computer vision [236]. During the last few years, pedestrian detection has made significant progress [256].

Li *et al.*. proposed "Scale-Aware Fast R-CNN (SAF R-CNN)" technique to detect pedestrian while achieving high accuracy on different dataset [Scale-Aware Fast R-CNN for Pedestrian Detection]. Liu *et al.* developed "CSP detector" for pedestrian detection where the authors achieved high accuracy on "CityPersons" and "Caltech" data-sets [238].

6) E-HEALTH SYSTEMS

Human detection can also play a substantial role in e-health systems, for instance, automatically elderly persons fall detection by activating an alarm. Since the number of senior citizens is rising significantly, fall detection has become more of a concern than ever before. Various devices and techniques have been proposed for fall detection [257]. Harrou *et al.* developed "MEWMA-based SVM" technique for fall detection while increasing the classification accuracy as well.

This model is also capable of categorizing the fall detection with high performance [242]. Considering illumination condition, Kong *et al.* came up with one technique to detect fall using RGB Depth camera [257]. Li *et al.* presented "Convolutional neural networks (CNNs)" based technique

TABLE 8. State-of-the-art human detection descriptors.

descriptor	classifier	train data	test data	Miss Rate	Accuracy
Informed Haar [180]	AdaBoost	Caltech	Caltech	34.60 %	–
Informed Haar [180]	AdaBoost	Inria	Inria	14.43 %	–
VJ [181]	AdaBoost	Inria	Inria	95%	–
VJ [54]	AdaBoost	Inria	Inria	72.48%	–
VJ [54]	AdaBoost	Inria	Caltech	94.73%	–
HOG [52]	linear SVM	Inria	MIT	68%	–
HOG [52]	linear SVM	Inria	Inria	45.98%	–
Shapelet [182]	AdaBoost	Inria	Caltech	91.37%	–
Shapelet [182]	AdaBoost	Inria	Inria	81.70%	–
MultiFtr+Motion [85]	linear SVM	TUD-Motion	Caltech	50.88%	–
MultiFtr+CSS [85]	AdaBoost	Inria	Caltech	60.89%	–
MultiFtr+CSS [85]	AdaBoost	Inria	Inria	24.74%	–
MB-BLP + WCRM [183]	EFLDA Classifier	Inria	Caltech	–	95%
MB-BLP+WCRM [183]	EFLDA Classifier	Inria	TUD-Brussels Pedestrian	–	96.53%
HogLbp [75]	linear SVM	Inria	Caltech	67.77%	–
HogLbp [75]	linear SVM	Inria	Inria	39.10%	–
LatSvm-V1 [184]	latent SVM	Pascal	Inria	43.83%	–
LatSvm-V1 [184]	latent SVM	Pascal	Caltech	79.78%	–
LatSvm-V2 [185]	latent SVM	Inria	Caltech	63.26%	–
LatSvm-V2 [185]	latent SVM	Inria	Inria	19.96%	–
ChnFtrs [186]	AdaBoost	Inria	Caltech	56.34%	–
ChnFtrs [186]	AdaBoost	Inria	Inria	22.18%	–
Shape [187]	K-means	Caltech	Inria	–	78.7%
Gradient distribution [188]	SVM	INRIA	INRIA	–	93.81%
Improved Shape Context [183]	SVM	OSU Infrared Image DB	OSU Infrared Image DB	–	90.54%
CrossTalk [189]	AdaBoost	Inria	Caltech	53.88%	–
CrossTalk [189]	AdaBoost	Inria	Inria	18.98%	–
VeryFast [190]	AdaBoost	Inria	Inria	15.96%	–
Roerei [190]	AdaBoost	Inria	Caltech	48.35%	–
Roerei [190]	AdaBoost	Inria	Inria	13.53%	–
AFS+Geo [191]	AdaBoost	Inria	Caltech	66.76%	–
FPDW [192]	AdaBoost	Inria	Inria	57%	–
PLS [193]	PLS+QDA	Inria	Inria	62%	–
POSEINV [194]	SVM	Inria	Inria	86%	–
Haar-like+HOG [195]	SVM	Inria	Inria	–	95%
HOG [195]	SVM	Inria	Inria	–	92.3%
ABM-HOG [196]	SVM	Inria	TUD-Pedestrian	–	90.2%
FeatSynth [197]	linear SVM	Inria	Caltech	60.16%	–
FeatSynth [197]	linear SVM	Inria	Inria	30.88%	–
MultiResC [198]	latent SVM	Caltech	Caltech	48.45%	–

TABLE 8. ***(Continued.)*** **State-of-the-art human detection descriptors.**

descriptor	classifier	train data	test data	Miss Rate	Accuracy
HikSvm [199]	HIK SVM	Inria	Inria	42.82%	–
HikSvm [199]	HIK SVM	Inria	Caltech	73.39%	–
SIFT [200]	Adaboost	Diamler	Diamler	34%	–
SIFT [200]	SVM	Diamler	Diamler	34%	–
SURF [200]	Adaboost	Diamler	Diamler	21%	–
SURF [200]	SVM	Diamler	Diamler	29%	–
SIFT [200]	Adaboost	Inria	Inria	49%	–
SIFT [200]	SVM	Inria	Inria	45%	–
SURF [200]	Adaboost	Inria	Inria	49%	–
SURF [200]	SVM	Inria	Inria	36%	–

to detect fall in video surveillance system while maintaining high accuracy [258].

7) HUMAN ACTIVITY RECOGNITION

Human activity recognition (HAR) has attracted considerable interest in the recent decade among the scientific community across the world. The ability of another person's action identification has become one of the key responsibilities in computer vision. Various techniques and devices have been advocated until now to recognize the human activities [294]. Chen *et al.* came up with "coordinate transformation and principal component analysis (CT-PCA) and online support vector machine (OSVM)" based technique for human activity recognition. The authors eliminated the orientation changes issue employing CT-PCA method [295]. Hassan *et al.* developed "smartphone inertial sensors-based" technique in which the authors integrated "kernel principal component analysis (KPCA), linear discriminant analysis (LDA)" and "Deep Belief Network (DBN)" to identify human actions [222].

Table 9 illustrates a brief evaluation of Abnormal Behavior, Crowd scene analysis, Person re-identification, Human tracking, Pedestrian detection, Elderly fall detection and Human activity recognition applications.

V. DATA-SET

A large number of human detection data-sets have been proposed in the recent decades and made them publicly accessible to check the performance of human detection techniques. We have collected these datasets based on various scenarios so that they can be applied in different applications as well. For instance, the CVC [37], Caltech [296] and ETH [38] datasets are compatible for pedestrian detection where as CAVIAR [297] and USC-B [298] are compatible for surveillance systems. Various datasets with their available Uniform Resource Locator (url), short descriptions and publications are listed (in TABLE 10).

VI. RESULT

Vision based human detection state-of-the-art methods have been investigated and denominated based on their outcomes; Log-Average Miss Rate (MR %) & Accuracy. The applied descriptors, classifiers, training and testing data-sets are reported (in TABLE 8).

In [180], Zhang *et al.* introduced the "Informed Haar" model, which avoided complete searches for each specific rectangle feature configuration and did not depend on random sampling. The outcome showed that this technique is stable in contradiction to occlusions and can perform at low computational cost. The authors used the model on two mostly familiar benchmark datasets named "INRIA" and "Caltech"; on "INRIA" it showed impressive result in regard to Log-Avg. Miss Rate.

In [54], Viola *et al.* described a human detection technique in which the authors integrated "Haar-like features" with "motion information" evaluated in a video series, taking into account two successive frames. The authors used face descriptor [181] to solve the issue of human detection but on "INRIA" and "Caltech" benchmark datasets, their method achieved high Log-Avg. Miss Rate.

In [52], Dalal *et al.* introduced the HoG features after studying the feature sets question for stable visual object identification. This method consists of vector space which calculates similarity by employing Euclidean or cosine distances while implementing a linear SVM. The experimental result of HoG outperformed existing human detection feature sets [195] Kim *et al.* proposed a human detection technique combining HoG features with Haar-like features whereby this model showed high accuracy compared to conventional HoG features. Li *et al.* presented a part-based human detection method [196] to overcome the pose changing and appearance shortcomings of human in complicated traffic regions wherein the authors applied a stochastic grammar technique. The authors designed the human appearance parts in an exuberant feature representation which increased the map

TABLE 9. A brief summary of Abnormal Behavior, Crowd scene analysis, Person re-identification, Human tracking, Pedestrian detection, Elderly fall detection and Human activity recognition applications.

methods	advantages	limitations
Abnormal Behavior	• Supervised methods show impressive performance for abnormal behaviors (labeled as classes) which are known to the machine [259]. • Abnormal behaviors can be detected at a high speed using semi-supervised methods [260], [261]. • Unsupervised methods are easy and can be used to quickly detect abnormal behaviors [262], [263]	• Supervised methods are useful to detect only particular behaviors, e.g., fighting [264], falling [13], and loitering [265]. • Supervised methods cannot detect ambiguous anomalies and is not practically feasible to learn all possible abnormal behaviors of humans [259]. • Semi-supervised methods are responsive to multiple parameters [266]. • Training time for unsupervised methods is relatively longer [262], [263]
Crowd scene analysis	• Real-time crowd analysis is feasible using continuum dynamics [267]. • Good performance of crowd scene analysis can be achieved even under lighting conditions [268]. • Short time occlusion problem can be resolved using matching-based tracking technique for crowd analysis [268].	• Present strategies generally target to analyze the crowd with high accuracy, without taking the computational time into account [269]. • Contextual knowledge for learning and tracking is not factored in the current crowd scene analysis techniques [269]. • If the view field is small with high crowd density and occlusion, then it becomes challenging to analyze the crowd scene [269].
Person re-identification	• Person re-identification is possible with intense occlusions employing fusing local-local and global-local matches [270], [271]. • Person re-identification is feasible even under illumination conditions and view changes [272], [273]	• Person re-identification techniques fail to cope up with the dynamic changes in the clothing patterns [274]. • Achieving a very high tracking accuracy by utilizing person re-identification techniques is not possible as of now [274].
Human tracking	• Human tracking is possible with appearance change which is important to achieve good performance [275]. • Real time human tracking is possible using online trackers [255], [276]. • Even in the presence of clutter backgrounds, human tracking is possible [30].	• It is still difficult to track human in real-time with offline trackers [255]. • Majority of the current works consider only two actions 'standing' or 'walking' while the other activities such as pose changes are ignored [277]. • Occlusion handling is a significant challenge which limits the accuracy of human tracking [277].
Pedestrian detection	• It is possible to detect pedestrian in real-time [191]. • Achieving a high efficiency in pedestrian detection is currently feasible [190]. • Partial occlusion can also be handled during pedestrian detection [180].	• Speed and accuracy are still low in real-time pedestrian detection [48]. • Occlusion, illumination conditions, view changes and noise still continue to remain as the open problems for pedestrian detection [38], [278], [279]
Elderly fall detection	• Elderly fall detection is possible at a low cost [280], [281]. • It is possible to detect elderly fall in less time while still maintaining a high accuracy [282], [283]	• Many existing strategies using vision-based detection lack flexibility for elderly fall detection [284]. • Sensor-based methods lack consistency in terms of providing a highly reliable and automatic fall detection systems [284].
Human activity recognition	• It is possible to recognize human action based on viewpoint changes using wearable sensors [285] and 3-D markers [286]. • Occluded actions can be recognized with the help of probabilistic and pose-based methods [287]–[291]	• As occluded part is impossible to extract, it is necessary to develop robust classifiers which can handle the occlusion even in dynamic backgrounds [292], [293]. • A significant barrier to achieve success in human action recognition is the absence of a benchmark dataset that efficiently represents camera motion [292], [293].

TABLE 10. Publicly accessible human detection data-sets.

name	publication	description	year	url
INRIA	[52], [299]–[301]	It contains 1208 humans in which 614 humans are for training, and 288 humans are for testing purpose.	2005	http://pascal.inrialpes.fr/data/human/
MIT	[302]	It contains 64 × 128 colour images, and 924 humans for training.	2000	http://cbcl.mit.edu/software-datasets/PedestrianData.html
Daimler-DB	[36], [303]	It contains colour images, for training 15.6k humans, 6.7k negative images and for testing 56.5k humans, 21.8k positive images.	2009	http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/Daimler_Mono_Ped__Detection_Be/daimler_mono_ped__detection_be.html
Daimler-CB	[304], [305]	It has 2.4k humans for training with 15k negative images and for testing 1.6k humans and 10k negative images.	2006	http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html
Penn–Fudan	[306]	This data-set contains 170 colour images with 345 humans.	2007	https://www.cis.upenn.edu/~jshi/ped_html/
H3D	[290]	It contains colour images in which 1500 humans for training and 500 humans for testing with 107 positive images.	2009	https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/h3d/
CVC	[37], [307]	It contains colour images, and for training, it has 2000 humans and 6175 negative images.	2007	http://www.cvc.uab.es/adas/
TUD-Brussels	[308]	It contains colour images in which 1776 humans, 218 negative images and 1098 positive images are for training while 1498 humans and 508 positive images are for testing.	2009	http://www.d2.mpi-inf.mpg.de/tud-brussels
TUD-det	[309]	It contains colour images in which 400 humans, 400 positive images are for training and 311 humans, 250 positive images are for testing.	2008	https://www.pgram.com/dataset/tud-pedestrians/
ETH	[38], [310]	It contains colour images in which 2388 humans are for training, and 12k humans are for testing.	2007	http://www.vision.ee.ethz.ch/aess/dataset/
USC-A	[298]	It contains grey images in which 205 positive images and 313 humans are for testing.	2005	http://www.iris.usc.edu/Vision-Users/OldUsers/bowu/DatasetWebpage/dataset.html
USC-B	[298]	It contains grey images in which 54 positive images and 271 humans are for testing.	2005	http://www.iris.usc.edu/Vision-Users/OldUsers/bowu/DatasetWebpage/dataset.html
USC-C	[311]	It contains grey images in which 100 positive images and 232 humans are for testing.	2007	http://www.iris.usc.edu/Vision-Users/OldUsers/bowu/DatasetWebpage/dataset.html

of HoG and ABM (Active Basis Model). Wang *et al.* combined HoG features with LBP (Local Binary Pattern) [75] for human detection to solve the partial occlusion issue; this method showed impressive performance on "INRIA" dataset.

In [182], Sabzmeydani *et al.* introduced a features set named "Shapelet" to solve the issue of recognizing humans in still scenes. These features were constructed of low-level features which consisted of gradient responses in scenes. Next, an "AdaBoost" classifier had been applied in a features

TABLE 10. ***(Continued.)*** **Publicly accessible human detection data-sets.**

name	publication	description	year	url
Caltech	[38], [296]	It contains colour images in which 192k humans with 61k negative images and 67k positive images for training. It has 155k humans with 56k negative images and 65k positive images for testing.	2009	https://dbcollection.readthedocs.io/en/latest/datasets/caltech_ped.html
MS COCO	[312]	It consists of 330K images with 2500k labelled instances. It has a total of 80 object classes, including human object.	2014	https://cocodataset.org/
ImageNet	[313]	The largest data-set is freely available on the internet that contains 14.19 million images with 20k categories.	2009	http://www.image-net.org/
PASCAL VOC	[22], [185], [314]	It contains 8,776 images with 20 classes.	2007	http://host.robots.ox.ac.uk/pascal/VOC/
EuroCity Persons	[315]	It has 47.3k colour images including 238.2k humans, annotations and bounding boxes are provided.	2019	https://eurocity-dataset.tudelft.nl/eval/user/login?_next=/eval/downloads/detection
PETA	[316]	It contains 19k colour images including 8705 humans, image resolution range starts from 17 × 39 to 169 × 365.	2014	http://mmlab.ie.cuhk.edu.hk/projects/PETA.htm
WiderPerson	[317]	It contains 13.4 colour images with 400K annotations, 8k images for training, 1k images for validation and 4.4k images for testing.	2019	http://www.cbsr.ia.ac.cn/users/sfzhang/WiderPerson/
Supervisely Person	[318], [319]	It contains 5.7k images with 6.9k annotated human instances.	2018	https://supervise.ly/explore/projects/supervisely-person-dataset-23304/datasets

set for learning purpose. Finally, "Shapelet" features had been used as input while training another classifier called "AdaBoost" to categorize between human and non-human. The experimental results showed that the model performed better on "INRIA" dataset.

In [85] Walk *et al.* executed motion features from optic flow which showed extensive performance on videos, surprisingly in the event of low-quality image sequences. Besides, the authors introduced a dynamic feature named "CSS" which is derived from self-similarity that continuously increases the performance of detection over various datasets not only in still images but in video sequences as well. Finally, the authors combined these two techniques with HoG features that surpassed until 20 percent of the state-of-the-art. "AdaBoost" and "Linear SVM" classifier were applied to categorize and analyze the performance on "Caltech", "TUD-Brussels" and "INRIA" datasets.

In [183], Chen *et al.* advocated a system to detect human in infrared images employing "improved Shape Context feature (ISC)" that is more stable to object distortion. The authors applied "Bilateral Filtering" for images in order to retain the extensive sharp edges information and to simplify the extracting task of the "Shape Context feature (SC)". The outcomes indicated that this method is apt to infrared images.

In [184], [185], Felzenszwalb *et al.* presented a human detection framework based on pictorial structures named "Deformable Part Model (DPM)". The authors used a scanning window technique that consisted of a comprehensive "root filter" and multiple "part filters". A latent variable "MI-SVM" construction called "latent SVM (LSVM)" is applied to train models using partly-labelled data. Deformable part-based methods obtained state-of-the-art achievement for object identification. However, it is not salient because of its dependency on heuristic initialization during training for the sake of the optimization of non-convex cost function.

In [186], Dollár *et al.* examined the "integral channel features" performance for object classification purposes, particularly human class. The "integral channel features" general concept is that various recorded image channels are calculated employing non-linear and linear input image transformations. After that, features such as ''local sums'', ''Haar features'', ''histograms'', and their various generalizations are computed accurately applying integral images. The authors demonstrated 'integral channel features' outperformed several features (for example, ''histogram of oriented gradient (HOG)'') if designed appropriately. The experimental works had been conducted mostly on ''INRIA'' dataset, and then the results had been shown on ''Caltech'' dataset as well. In [189], Dollár *et al.* presented two opposing frameworks: ''promising neighbors excitation'' and ''inferior neighbors inhibition'' detectors to communicate with ''crosstalk cascades''. Subsequently, the authors introduced an optimized application integrated with current approaches in ''fast multi-scale feature'' computation which achieved state-of-the-art accuracy.

In [187], Li *et al.* advocated an improved ''Implicit Shape Model (ISM)'' in which the authors examined six familiar key point detectors to find out the best one to detect human and on-road traffic. The experimental result showed that ''Harris Detector'' was the best detector. Then, a fuzzy function was used to overcome the original ''shape context local feature'' detector's shortcomings. Lastly, the authors applied ''Mean shift'' instead of ''k-means algorithm'' for codebook creation accurately on datasets.

In [188], Mehralian *et al.* developed a dynamic method based on sliding window to detect human in static images and image sequences. All windows were split into overlapping cells. From these cells, features were extricated which were derived from gradient distribution analysis. The authors applied ''Principal component analysis (PCA)'' on a cell computation to get the cell features. Finally, ''Support Vector Machine (SVM)'' was used to classify the features and tested on ''INRIA'' dataset. The presented model is more stable in noise and showed impressive detection accuracy compared with ''Histograms of Oriented Gradient (HOG)'' method. Benenson *et al.* [190] presented a novel human descriptor which enhanced not only quality but also speed by handling various computation of scales and transferring successfully over the state-of-the-art. This model provided an excellent detection rate on monocular images. In addition, the authors developed a dynamic technique for utilizing ''geometric context'' retrieved from stereo images.

In [191], Levi *et al.* addressed part-based ''Accelerated Feature Synthesis (AFS)'' technique that used various tactics for decreasing the searching locations number for all parts. The authors developed the ''KDFerns'' method to compare all locations of the image with the model parts subset only. By applying ''spatial inhibition'' and an object-level ''coarse-to-fine'' technique, candidate part positions for a particular part were additionally decreased. This method achieved real time running efficiency while maintaining the same accuracy of the main ''Feature Synthesis'' on ''INRIA'' and ''Caltech'' datasets. Meanwhile, Schwartz *et al.* described an ''edge-based features'' [193] technique in which the authors applied ''Partial Least Squares (PLS)'' analysis that outperformed the state-of-the-art methods on ''INRIA'', ''DaimlerChrysler'', and ''ETHZ pedestrian'' datasets. Park *et al.* [198] presented a multiresolution technique that behaved like a ''deformable part-based'' method. The authors applied this model on ''Caltech'' dataset and reduced the detection missed rate to 29 percent.

In [194], Lin *et al.* presented a learning-based ''sliding window-style'' global descriptor to learn and to categorize the image patterns of human/non-human by splitting up the human poses and shapes simultaneously, and removing articulation-nonreactive features. The authors used ''Histograms of oriented gradients (HoG)'' as a low-level feature source and ''kernel SVMs'' for classification. The proposed method showed impressive result on ''INRIA'' dataset. Besides, ''MIT-CBCL pedestrian'' dataset was also used to evaluate the performance of the methods. Later, in [197], the authors introduced a learning part-based method which included an iterative feature generation and pruning process. The authors used ''Predictive Feature Selection (PFS)'' with ''linear SVM'' for feature pruning. This model was evaluated on ''INRIA'', ''Caltech'' and another customized datasets where it made progress the state-of-the-art on ''Caltech'' dataset.

In [199], Maji *et al.* introduced a ''multi-level histograms of oriented edge energy'' based dynamic features in which the authors used ''intersection kernel SVMs (IKSVMs)'' for classification. The authors applied the method on various datasets wherein this method showed impressive result on ''INRIA'' dataset and improved the detection speed on ''Caltech'' dataset. In [200], Yao *et al.* compared various detectors including ''SIFT'' & ''SURF'' descriptors for human detection in which the authors used ''AdaBoost'' and ''SVM'' for classification. The authors applied these descriptors on ''Diamler'', and ''INRIA'' datasets and the experiment showed significant result on ''INRIA'' dataset.

VII. SUGGESTIONS, OPEN PROBLEMS AND FUTURE DIRECTIONS

In the subsequent sections, various suggestions are given to improve the described feature extraction techniques in terms of speed and accuracy. Various open problems and future directions are also provided for those who are keen to work in the area of human detection.

A. SUGGESTIONS

Actual object detection research started in 2001 after the discovery of the Viola-Jones technique. Several research articles have been published since to improve the Viola-Jones algorithm. For any object detection, accuracy is the most important factor while maintaining the speed. If linear SVM model [320] is to applied to predict feature image in training dataset and change the scaling factor using genetic algorithm [321], the efficiency of Viola-Jones algorithm can be increased.

The structure of Scale Invariant Feature Transform (SIFT) is complicated and its computational time is high. To increase its efficiency and reduce its computational time a hybrid method has been proposed. First, the linear combination of cityblock distance and chessboard distance should be used instead of euclidean distance to increase the SIFT feature matching efficiency. Besides, character point should be reduced while evaluating part-feature results [322]. In the next step, SIFT descriptor should be improved with the polar histogram orientation bin. Here, rectangular region should be replaced with a circular region because of its' greater invariance in rotation [323]. Finally, to achieve higher accuracy a fuzzy closed-loop control technique should be employed. In open-loop techniques, the outcome of each step relies on the preceding step. Hence, errors are accrued on the whole recognition system and generated to the ultimate step. Thus, the end result appears to be vulnerable to error and is unreliable. This issue can be solved by applying a fuzzy control technique. The approach is non-linear and there is no mathematical model available [324].

In order to increase the accuracy and speed of SURF algorithm, a new method has been proposed which is combined with the conventional SURF technique. At the beginning step, SURF (Speeded-Up Robust Features) combined with FAST method should be used to detect and describe the feature points of images. Laplace operator should be employed on weighted FAST feature points to get the quick matching result on the target extraction. For this operation, another feature point extraction will be secured. As a result, SURF descriptors will get more robustness in a matter of quick matching [325]. Later, RANdom SAmple Consensus (RANSAC) should be employed to eliminate the mismatch pairs. RANSAC algorithm is more effective and robust compared to other estimation methods [326].

Numerous methods have been developed until recently to improve the performance of Bag of Words model. A dynamic technique has been suggested to increase the accuracy for object categorization based on Bag of Words algorithm. First of all, Difference-of-Gaussian (DoG) will be employed to detect the interest points as key point detection is one of the most important steps for extracting the features of BoW model [327]. Secondly, SIFT feature should be replaced with PCA-SIFT to describe the key points of the local descriptor. This is because PCA (principal component analysis) reduces the dimensionality of SIFT descriptor. Therefore, not only image retrieval performance will be increased but matching will be faster as well. Next, vector quantization clustering algorithm k-means should be replaced by random forest method to overcome the limitation of k-means. By applying random forest clustering, accuracy will be increased and computational cost will be reduced [328]. Finally, for the purpose of image annotation, classification algorithm-linear-SVM should be applied because of its efficiency and robustness [329].

Since HoG (Histograms of Oriented Gradients) is one of the most popular object detection descriptors in computer vision, many studies based on HoG algorithm have been carried out to improve its techniques. After comparing and analyzing many related studies, a novel hybrid method has been proposed to increase its accuracy and speed. In the first stage, PCA (Principal Component Analysis) and HoG will be combined together. PCA is employed to HoG features for getting the score vectors (PCA-HoG). PCA reduces the dimensionality of HoG descriptor; consequently, original descriptor will become more robust in object detection. Next, SFS (Stepwise Forward Selection) or SBS (Stepwise Backward Selection) technique will be applied to select an appropriate part of PCA-HoG element vectors. Then, classifier will use these PCA-HoG element vectors as an input to categorize the particular object [330]. Finally, SVM classifier will be replaced by GA-XBoost [331] to increase the classification accuracy.

Among all the computer vision algorithms, Deformable Part Model shows the highest detection accuracy. In contrast to other improved techniques, DPM is slower in speed. Consequently, a new dynamic technique has been recommended to make DPM faster and more accurate. At the initial stage, we recommend replacing the HoG descriptor with WHoG. Weighted Histograms of Oriented Gradients (WHoG) [73] is a combination of global shape descriptors and local point descriptors which describe object detection with diverse textures. Bottom-up pose clustering method is used to handle intense pose variations. Two CNNs work parallel there; one is input pose taker templates and the other is taker of color information. First, the authors explain pose clustering technique and employ an improved HoG descriptor (WHoG) to make a pose-specific bird detector. WHoG generates a result for every candidate posture, with the maximum score indicating the recognized position for that specific posture. In the subsequent step, scale invariant color components have been used to make a spices-specific object detector. WHoG finds out the object and defines its position exactly in any particular picture by allocating additional weight to fringe and less load to body structures or stripes. WHoG can be used to detect any kind of object that conveys intestinal designs or textures. This approach can detect textured objects in opposition to background clutter. In addition, it has an immense dimensional articulation as well as pose variety such as small objects like birds. Integration of properly schematic scale invariant color features into the method increases the detection accuracy. This method minimizes computational complication and shows a great performance progress on a comprehensive dataset: CUB-200-2011. Finally, WHoG features will be used to train the latent SVM (LSVM) classifier.

B. OPEN PROBLEMS AND FUTURE DIRECTIONS

1) OPEN-WORLD LEARNING AND ACTIVE VISION

Rapid changes in fashion trends and latest inventions require the detection systems to be continuously modified, introducing new classes, or upgrading existing ones in order to identify objects more effectively and accurately. Following an unsupervised way, new classifiers can be built using the existing one without any extra effort of learning new object classes.

2) MULTI-MODAL DETECTION

New sensory modalities have been developed very recently, especially depth and thermal cameras [332], [333]. The methods used for visual images are, however, often used for thermal images and to a lesser extent for profound depth images. While thermal images help differentiate the foreground from the background, this can only be used for infrared light irradiating artifacts (e.g., mammals, heating). It is simple to segment objects using depth images but generic methods have not been proposed for detecting particular classes (i.e., human class) as higher resolution images are required to do so. Depth and thermal cameras alone do not seem sufficient for object detection, at least with their current resolution, but progress can be expected with the advancements in sensing technology.

3) OBJECT-PART RELATION

Detecting the object first or the parts first is a very basic dilemma during the detection process with no exact solution. This search of object and its parts needs to be carried out simultaneously where each of these provides feedback to each other. Doing this still remains a great problem which relates to the usage of context information. Furthermore, in the case of the object component, the interaction of many hierarchies can also be broken down into sub-parts, and what should be done first is not clear in general.

4) PIXEL-LEVEL DETECTION (SEGMENTATION) AND BACKGROUND OBJECTS

Successful detection of all objects in a single scene along with a proper understanding of the scene will require pixel level detection of the objects and a 3D model of the scene where most detectors till date use 2D images. Accordingly, object detection and image segmentation require integration at some point. Achieving this automatic worldview is still far from reality and active vision mechanisms could be important to achieve this [334], [335].

5) OCCLUSION

Although relevant research has been carried out by [298], partial occlusion remains an influential problem, but no solution exists as of now. For this type of problem, the deformable part-based model [185] has succeeded to a certain extent. However, further improvement in performance remains an open problem.

6) MULTI-VIEW, MULTI-POSE, MULTI-RESOLUTION

Detection of human object have been designed by various methods under a single view. Among these methods, only deformable part-based models are able to deal with particular pose variations, whereas other methods are unable to handle multiple view and large pose variations.

7) EFFICIENCY

Efficiency is a prime concern in any object detection system. It does not imply real-time performance, and works for instance, deformable part-based model [185] is efficient and robust, yet not quick enough for real-time solutions. Nonetheless, the usage of specialized software such as GPU provides real-time run for some methods like deep learning.

8) DETECTOR SELECTION

With so many options in feature representation and classifier, selecting the best option is not an easy choice. In this article, we have presented various techniques with their successes and shortcomings. We have provided the options in which to select techniques which perform the best on a particular data-set. The major advantage of vision-based technique is that it does not need a huge data-set like deep learning. Recently, there is a big hype regarding deep learning – with valid reasons such as high accuracy. We do not have to handcraft design the features; rather they will automatically learn so. However, it requires massive data to train and prepare computational resources. Due to its semi-blackbox nature, debugging can be challenging. Combination of both machine-learning and deep learning can be very useful where researchers can use deep learning to extract the best features and then researchers can choose a machine-learning classifier that produces the best result of a particular data-set.

VIII. CONCLUSION

In the most recent decade, human detection topic has gained impressive attention among the research community due to its large number of applications. This article provided a comprehensive review on the state-of-the-art feature extraction techniques followed by various classifiers. It also critically analyzed the various techniques by explaining the possible pros and cons in the light of the application scenarios. The distinguishing characteristic of the article is that it presented a more detailed evaluation and analysis of the existing feature extraction techniques with their invariants, which were otherwise not included in the currently available literature works, with respect to various performance indicators, such as log-rate and accuracy. Various publicly available datasets for the feature extraction techniques for human detection were reported in conjunction with a concise description. Although techniques such as Viola-Jones and SURF can detect objects in real-time and overcome sift limitations, they are still sensitive to illuminated conditions. Other techniques such as SIFT, BoW, and OMs do exist and provide other interesting benefits which include insensitivity to occlusion and clutters, simplicity, and low-order element construction, but they are expensive from a computational standpoint. HoG is yet another technique that possesses some interesting features such as its invariance towards photometric and geometric transformations and illuminated conditions but faces a setback as it misses the context of spatially neighboring pixels.

Upon thoroughly reviewing all the existing feature extraction techniques, it was concluded that the DPM technique performs relatively better than its counterparts primarily because it can provide an optimum performance across multiple aspects such as the ability to handle particular pose variations, multiple views, and is application-free i.e., works reasonably well in real-time systems. However, the technique also poses certain limitation which deserve additional attention and scrutiny. The DPM technique is compute-intensive since it depends on the heuristic initialization during the training process in order to optimize the non-convex cost function. Despite this, the DPM technique is still relatively superior compared to the other existing techniques for human detection which can find fundamental importance in many real-time applications involving human detection systems. This article is not only for the researchers in the area of computer vision but also intended for people who are keen on delving into the area of human detection utilizing machine learning algorithms. As presented in the preceding section, possible future research directions includes but not limited to further enhancement of the speed, computational training time, and the accuracy of algorithms.

REFERENCES

[1] M. Cohen-McFarlane, R. Goubran, and F. Knoefel, "Novel coronavirus cough database: NoCoCoDa," *IEEE Access*, vol. 8, pp. 154087–154094, 2020.

[2] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. Y. Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," *Science*, vol. 368, no. 6489, pp. 395–400, Apr. 2020.

[3] L. Zhong, L. Mu, J. Li, J. Wang, Z. Yin, and D. Liu, "Early prediction of the 2019 novel coronavirus outbreak in the mainland China based on simple mathematical model," *IEEE Access*, vol. 8, pp. 51761–51769, 2020.

[4] F. Amanat, D. Stadlbauer, S. Strohmeier, T. H. O. Nguyen, V. Chromikova, M. McMahon, K. Jiang, G. A. Arunkumar, D. Jurczyszak, J. Polanco, and M. Bermudez-Gonzalez, "A serological assay to detect SARS-CoV-2 seroconversion in humans," *Nature Med.*, vol. 26, pp. 1–4, Jul. 2020.

[5] (2021). *Coronavirus Update (Live)*. Accessed: Feb. 22, 2021. [Online]. Available: https://www.worldometers.info/coronavirus

[6] (2020). *There is a Current Outbreak Coronavirus (COVID-19) Disease*. Accessed: Sep. 2, 2020. [Online]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_1

[7] H. Ritchie, "Natural disasters," in *Our World in Data*. 2019. Accessed: Aug. 25, 2020. [Online]. Available: https://ourworldindata.org/natural-disasters

[8] J. Yeung. (2020). *Australia's Deadly Wildfires are Showing no Signs of Stopping. Here's What You Need to Know*. CNN. Accessed: Aug. 25, 2020. [Online]. Available: https://edition.cnn.com/2020/01/01/australia/australia-fires-explainer-intl-hnk-scli/index.html

[9] M. Shah. (2020). *Official Number of New Zealand Volcano Victims Rises to 18 After Death in Australia*. Global News. Accessed: Aug. 25, 2020. [Online]. Available: https://globalnews.ca/news/6402259/new-zealand-volcano-death-toll-18/

[10] M. F. Sohail, C. Y. Leow, and S. Won, "Non-orthogonal multiple access for unmanned aerial vehicle assisted communication," *IEEE Access*, vol. 6, pp. 22716–22727, 2018.

[11] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-depth sensors for fall detection: A review," *IEEE Sensors J.*, vol. 17, no. 12, pp. 3585–3604, Jun. 2017.

[12] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, and N. Noury, "Complete sound and speech recognition system for health smart homes: Application to the recognition of activities of daily living," New Develop. Biomed. Eng., In-Tech, Tech. Rep., 2010, pp. 645–673.

[13] E. E. Stone and M. Skubic, "Fall detection in homes of older adults using the Microsoft Kinect," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 1, pp. 290–301, Jan. 2015.

[14] (2020). *Road Traffic Injuries*. Accessed: Sep. 2, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[15] (2020). *Statistics of Road Traffic Accidents in Europe and North America*. Accessed: Sep. 2, 2020. [Online]. Available: https://www.unece.org/info/media/news/statistics/2016/statistics-of-road-traffic-accidents-in-europe-and-north-america/doc.html

[16] Y. Imamura, S. Okamoto, and J. H. Lee, "Human tracking by a multi-rotor drone using HOG features and linear SVM on images captured by a monocular camera," in *Proc. Int. MultiConference Eng. Comput. Scientists*, vol. 1, 2016, pp. 8–13.

[17] K. Suresh, V. Jeoti, M. Drieberg, and A. Iqbal, "On self driving cars: An LED time of flight (ToF) based detection and ranging using various unipolar optical CDMA codes," in *Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jun. 2019, pp. 1–6.

[18] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, and P. Maragos, "A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1169–1173.

[19] C. A. H. Ritchie, J. Hasell, and M. Roser, "Terrorism," in *Our World Data*. 2019. Accessed: Aug. 25, 2020. [Online]. Available: https://ourworldindata.org/terrorism

[20] M. Alsharif. (2020). *New York 9/11 Victim Identified 18 Years After Attack*. CNN. Accessed: Aug. 25, 2020. [Online]. Available: https://edition.cnn.com/2019/10/18/us/9-11-victim-identified-18-years-later/index.html

[21] H. Agerholm. (2020). *Oxford Street Incident: Two Men Hand Themselves Into Police After 'fight' That Triggered Panic*. The Independent. Accessed: Aug. 25, 2020. [Online]. Available: https://www.independent.co.uk/news/uk/home-news/oxford-street-fight-triggered-terror-evacuation-british-transport-police-a8075591.html

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[23] B. M. Mahmmod, A. M. Abdul-Hadi, S. H. Abdulhussain, and A. Hussien, "On computational aspects of krawtchouk polynomials for high orders," *J. Imag.*, vol. 6, no. 8, p. 81, Aug. 2020.

[24] S. S. Sumit, J. Watada, A. Roy, and D. Rambli, "In object detection deep learning methods, YOLO shows supremum to Mask R-CNN," *J. Phys., Conf. Ser.*, vol. 1529, no. 4, 2020, Art. no. 042086.

[25] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[26] H. Idrees, K. Soomro, and M. Shah, "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1986–1998, Oct. 2015.

[27] J. Steffens, E. Elagin, and H. Neven, "PersonSpotter-fast and robust system for human detection, tracking and recognition," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 516–521.

[28] S. S. Ghidary, Y. Nakata, T. Takamori, and M. Hattori, "Human detection and localization at indoor environment by home robot," in *Proc. Conf. IEEE Int. Conf. Syst., Man Cybern. (SMC), Cybern. Evolving Syst., Hum., Org., Complex Interact.*, vol. 2, Oct. 2000, pp. 1360–1365.

[29] S. S. Ghidary, Y. Nakata, T. Takamori, and M. Hattori, "Localization and approaching to the human by mobile home robot," in *Proc. 9th IEEE Int. Workshop Robot Hum. Interact. Commun. (IEEE RO-MAN)*, Sep. 2000, pp. 63–68.

[30] J. Zhou and J. Hoang, "Real time robust human detection and tracking system," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Sep. 2005, p. 149.

[31] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1631–1643, Oct. 2005.

[32] R. Furman and J. T. Kinn, *Practical Tips for Publishing Scholarly Articles: Writing and Publishing in the Helping Professions*. Washington, DC, USA: American Psychological Association, 2012.

[33] J. D. Culler and K. Lamb, *Just Being Difficult?: Academic Writing in the Public Arena*. Palo Alto, CA, USA: Stanford Univ. Press, 2003.

[34] A. Angeleas, N. Bourbakis, and G. Tsihrintzis, "Categorization of research surveys and reviews on human activities," in *Proc. 7th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2016, pp. 1–6.

[35] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sep. 2007.

[36] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[37] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[38] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2011.

[39] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 613–627.

[40] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019.

[41] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognit.*, vol. 36, no. 3, pp. 585–601, Mar. 2003.

[42] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Comput. Vis. Image Understand.*, vol. 81, no. 3, pp. 231–268, Mar. 2001.

[43] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, Nov. 2006.

[44] H. Zhou and H. Hu, "Human motion tracking for rehabilitation—A survey," *Biomed. Signal Process. Control*, vol. 3, no. 1, pp. 1–18, Jan. 2008.

[45] B. Schiele, M. Andriluka, N. Majer, S. Roth, and C. Wojek, "Visual people detection: Different models, comparison and discussion," in *Proc. IEEE ICRA Workshop People Detection Tracking*, May 2009, pp. 1–8.

[46] M. Paul, S. M. Haque, and S. Chakraborty, "Human detection in surveillance videos and its applications—A review," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, p. 176, 2013.

[47] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognit.*, vol. 51, pp. 148–175, Mar. 2016.

[48] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.

[49] T. Santhanam, C. P. Sumathi, and S. Gomathi, "A survey of techniques for human detection in static images," in *Proc. 2nd Int. Conf. Comput. Sci., Eng. Inf. Technol. (CCSEIT)*, 2012, pp. 328–336.

[50] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, no. 2, Sep. 1999, pp. 1150–1157.

[51] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 404–417.

[52] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Washington, DC, USA: IEEE Computer Society, May 2005, pp. 886–893.

[53] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, pp. 511–518.

[54] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, Jul. 2005.

[55] A. Agarwal, S. Gupta, and D. K. Singh, "Review of optical flow technique for moving object detection," in *Proc. 2nd Int. Conf. Contemp. Comput. Informat. (IC I)*, Dec. 2016, pp. 409–413.

[56] S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 211–223, Feb. 2016.

[57] J. Hernández-Aceituno, L. Acosta, and J. D. Piñeiro, "Pedestrian detection in crowded environments through Bayesian prediction of sequential probability matrices," *J. Sensors*, vol. 2016, pp. 1–8, Jan. 2016.

[58] C. Geng and X. Jiang, "Face recognition using SIFT features," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 3313–3316.

[59] W.-S. Lin, Y.-L. Wu, W.-C. Hung, and C.-Y. Tang, "A study of real-time hand gesture recognition using SIFT on binary images," in *Proc. Adv. Intell. Syst. Appl.*, vol. 2. Berlin, Germany: Springer, 2013, pp. 235–246.

[60] X. Hu, Y. Tang, and Z. Zhang, "Video object matching based on SIFT algorithm," in *Proc. Int. Conf. Neural Netw. Signal Process.*, Jun. 2008, pp. 412–415.

[61] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using SIFT features and mean shift," *Comput. Vis. Image Understand.*, vol. 113, no. 3, pp. 345–352, Mar. 2009.

[62] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[63] L.-C. Chiu, T.-S. Chang, J.-Y. Chen, and N. Y.-C. Chang, "Fast SIFT design for real-time visual feature extraction," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3158–3167, Aug. 2013.

[64] Y. Sakai, T. Oda, M. Ikeda, and L. Barolli, "An object tracking system based on SIFT and SURF feature extraction methods," in *Proc. 18th Int. Conf. Netw.-Based Inf. Syst.*, Sep. 2015, pp. 561–565.

[65] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, "A comparative study of SIFT and its variants," *Meas. Sci. Rev.*, vol. 13, no. 3, pp. 122–131, Jun. 2013.

[66] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 4, Jun. 2004, pp. 506–513.

[67] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, nos. 10–12, pp. 1771–1787, Jun. 2008.

[68] V. Seib, M. Kusenbach, S. Thierfelder, and D. Paulus, "Object recognition using Hough-transform clustering of SURF features," in *Proc. Workshops Electronical Comput. Eng. Subfields*, 2014, pp. 169–176.

[69] B. Anand and P. K. Shah, "Face recognition using SURF features and SVM classifier," *Int. J. Electron. Eng. Res.*, vol. 8, no. 1, pp. 1–8, 2016.

[70] D. Schmitt and N. McCoy, "Object classification and localization using SURF descriptors," *CS*, vol. 229, pp. 1–5, Dec. 2011.

[71] J. Canny, "A computational approach to edge detection," *Readings in Computer Vision*. Amsterdam, The Netherlands: Elsevier, 1987, pp. 184–203.

[72] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[73] D. Karmaker, I. Schiffner, R. Strydom, and M. V. Srinivasan, "WHoG: A weighted HoG-based scheme for the detection of birds and identification of their poses in natural environments," in *Proc. 14th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2016, pp. 1–7.

[74] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. Int. Conf. Comput. Vis.*, vol. 1, no. 2, Nov. 2011, p. 6.

[75] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 32–39.

[76] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[77] S. Yao, S. Pan, T. Wang, C. Zheng, W. Shen, and Y. Chong, "A new pedestrian detection method based on combined HOG and LSS features," *Neurocomputing*, vol. 151, pp. 1006–1014, Mar. 2015.

[78] A. K. Sah, S. Bhowmik, S. Malakar, R. Sarkar, E. Kavallieratou, and N. Vasilopoulos, "Text and non-text recognition using modified HOG descriptor," in *Proc. IEEE Calcutta Conf. (CALCON)*, Dec. 2017, pp. 64–68.

[79] N. Chen, W.-N. Chen, and J. Zhang, "Fast detection of human using differential evolution," *Signal Process.*, vol. 110, pp. 155–163, May 2015.

[80] M. Villamizar, J. Scandaliaris, A. Sanfeliu, and J. Andrade-Cetto, "Combining color-based invariant gradient detector with HoG descriptors for robust image detection in scenes under cast shadows," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 1997–2002.

[81] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, and C. L. Tan, "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognit.*, vol. 51, pp. 125–134, Mar. 2016.

[82] M. Ghorbani, A. T. Targhi, and M. M. Dehshibi, "HOG and LBP: Towards a robust face recognition system," in *Proc. 10th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Oct. 2015, pp. 138–141.

[83] H. Lahiani and M. Neji, "Hand gesture recognition method based on HOG-LBP features for mobile devices," *Procedia Comput. Sci.*, vol. 126, pp. 254–263, Jan. 2018.

[84] G. Gan and J. Cheng, "Pedestrian detection based on HOG-LBP feature," in *Proc. 7th Int. Conf. Comput. Intell. Secur.* New York, NY, USA: Springer, Dec. 2011, pp. 715–720.

[85] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1030–1037.

[86] Y. Goto, Y. Yamauchi, and H. Fujiyoshi, "CS-HOG: Color similarity-based HOG," in *Proc. 19th Korea-Japan Joint Workshop Frontiers Comput. Vis.*, Jan. 2013, pp. 266–271.

[87] D. L. Cosmo, E. O. T. Salles, and P. M. Ciarelli, "Pedestrian detection system based on HOG and a modified version of CSS," in *Proc. 7th Int. Conf. Mach. Vis. (ICMV)*. Washington, DC, USA: International Society for Optics and Photonics, vol. 9445, Feb. 2015, p. 94450.

[88] Y. Pang, Y. Yuan, X. Li, and J. Pan, "Efficient HOG human detection," *Signal Process.*, vol. 91, no. 4, pp. 773–781, Apr. 2011.

[89] Y. Pang, H. Yan, Y. Yuan, and K. Wang, "Robust CoHOG feature extraction in human-centered image/video management system," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 458–468, Apr. 2012.

[90] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 141–154, Feb. 2016.

[91] J. Liu, G. Zhang, Y. Liu, L. Tian, and Y. Q. Chen, "An ultra-fast human detection method for color-depth camera," *J. Vis. Commun. Image Represent.*, vol. 31, pp. 177–185, Aug. 2015.

[92] C. Liu, X. Lu, S. Ji, and W. Geng, "A fog level detection method based on image HSV color histogram," in *Proc. IEEE Int. Conf. Prog. Informat. Comput.*, May 2014, pp. 373–377.

[93] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[94] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 524–531.

[95] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 985–992.

[96] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate Web image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 25–32.

[97] K.-T. Chen, K.-H. Lin, Y.-H. Kuo, Y.-L. Wu, and W. H. Hsu, "Boosting image object retrieval and indexing by automatically discovered pseudo-objects," *J. Vis. Commun. Image Represent.*, vol. 21, no. 8, pp. 815–825, Nov. 2010.

[98] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2169–2178.

[99] M. Shahiduzzaman, D. Zhang, and G. Lu, "Improved spatial pyramid matching for image classification," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 449–459.

[100] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[101] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 1–15.

[102] J. Farooq, "Object detection and identification using SURF and BoW model," in *Proc. Int. Conf. Comput., Electron. Electr. Eng. (ICE Cube)*, Apr. 2016, pp. 318–323.

[103] N. M. Ali, S. W. Jun, M. S. Karis, M. M. Ghazaly, and M. S. M. Aras, "Object classification and recognition using bag-of-words (BoW) model," in *Proc. IEEE 12th Int. Colloq. Signal Process. Appl. (CSPA)*, Mar. 2016, pp. 216–220.

[104] D. Yan, K. Li, S. Gu, and L. Yang, "Network-based bag-of-words model for text classification," *IEEE Access*, vol. 8, pp. 82641–82652, 2020.

[105] L. Zhu, H. Jin, R. Zheng, and X. Feng, "Weighting scheme for image retrieval based on bag-of-visual-words," *IET Image Process.*, vol. 8, no. 9, pp. 509–518, Sep. 2014.

[106] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.

[107] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[108] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Int. Workshop Workshop Multimedia Inf. Retr. (MIR)*, 2007, pp. 197–206.

[109] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th Int. Conf. Adv. Geographic Inf. Syst. (GIS SIGSPATIAL)*, 2010, pp. 270–279.

[110] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010.

[111] G. Csurka and F. Perronnin, "Fisher vectors: Beyond bag-of-visual-words image representations," in *Proc. Int. Conf. Comput. Vis., Imag. Comput. Graph.* Berlin, Germany: Springer, 2010, pp. 28–42.

[112] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[113] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 836–849.

[114] P. Ott and M. Everingham, "Shared parts for deformable part-based models," in *Proc. CVPR*, Jun. 2011, pp. 1513–1520.

[115] Q. Zhou, G. Wang, K. Jia, and Q. Zhao, "Learning to share latent tasks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2264–2271.

[116] S. Branson, G. Van Horn, C. Wah, P. Perona, and S. Belongie, "The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization," *Int. J. Comput. Vis.*, vol. 108, nos. 1–2, pp. 3–29, Feb. 2014.

[117] X. Chen and A. L. Yuille, "Parsing occluded people by flexible compositions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3945–3954.

[118] G. Ghiasi, Y. Yang, D. Ramanan, and C. C. Fowlkes, "Parsing occluded people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2401–2408.

[119] J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2497–2504.

[120] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3124–3129.

[121] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[122] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. CVPR*, Jun. 2011, pp. 1385–1392.

[123] E. Trulls, S. Tsogkas, I. Kokkinos, A. Sanfeliu, and F. Moreno-Noguer, "Segmentation-aware deformable part models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 168–175.

[124] R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester, "Object detection with grammar models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 442–450.

[125] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[126] J. Li, Y. Yin, X. Liu, D. Xu, and Q. Gu, "12,000-fps multi-object detection using HOG descriptor and SVM classifier," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5928–5933.

[127] J. Li, X. Liu, F. Liu, D. Xu, Q. Gu, and I. Ishii, "A hardware-oriented algorithm for ultra-high-speed object detection," *IEEE Sensors J.*, vol. 19, no. 10, pp. 3818–3831, May 2019.

[128] C.-Y. Wee, R. Paramesran, R. Mukundan, and X. Jiang, "Image quality assessment by discrete orthogonal moments," *Pattern Recognit.*, vol. 43, no. 12, pp. 4055–4068, Dec. 2010.

[129] B. Xiao, L. Li, Y. Li, W. Li, and G. Wang, "Image analysis by fractional-order orthogonal moments," *Inf. Sci.*, vols. 382–383, pp. 135–149, Mar. 2017.

[130] P.-T. Yap, R. Paramesran, and S.-H. Ong, "Image analysis by Krawtchouk moments," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1367–1377, Nov. 2003.

[131] R. Mukundan, S. H. Ong, and P. A. Lee, "Image analysis by Tchebichef moments," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1357–1364, Sep. 2001.

[132] S. H. Abdulhussain, A. R. Ramli, S. A. R. Al-Haddad, B. M. Mahmmod, and W. A. Jassim, "On computational aspects of Tchebichef polynomials for higher polynomial order," *IEEE Access*, vol. 5, pp. 2470–2478, 2017.

[133] H. Karmouni, T. Jahid, I. E. Affar, M. Sayyouri, A. Hmimid, H. Qjidaa, and A. Rezzouk, "Image analysis using separable Krawtchouk-Tchebichef's moments," in *Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2017, pp. 1–5.

[134] H. Zhu, M. Liu, H. Shu, H. Zhang, and L. Luo, "General form for obtaining discrete orthogonal moments," *IET Image Process.*, vol. 4, no. 5, pp. 335–352, Oct. 2010.

[135] S. H. Abdulhussain, A. R. Ramli, B. M. Mahmmod, M. I. Saripan, S. A. R. Al-Haddad, T. Baker, W. N. Flayyih, and W. A. Jassim, "A fast feature extraction algorithm for image and video processing," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[136] A. M. Abdul-Hadi, S. H. Abdulhussain, and B. M. Mahmmod, "On the computational aspects of Charlier polynomials," *Cogent Eng.*, vol. 7, no. 1, Jan. 2020, Art. no. 1763553.

[137] R. Koekoek, P. A. Lesky, and R. F. Swarttouw, *Hypergeometric Orthogonal Polynomials and Their Q-Analogues*. Berlin, Germany: Springer, 2010.

[138] H. Karmouni, A. Hmimid, T. Jahid, M. Sayyouri, H. Qjidaa, and A. Rezzouk, "Fast and stable computation of the charlier moments and their inverses using digital filters and image block representation," *Circuits, Syst., Signal Process.*, vol. 37, no. 9, pp. 4015–4033, Sep. 2018.

[139] T. Jahid, H. Karmouni, A. Hmimid, M. Sayyouri, and H. Qjidaa, "Fast computation of Charlier moments and its inverses using Clenshaw's recurrence formula for image analysis," *Multimedia Tools Appl.*, vol. 78, no. 9, pp. 12183–12201, May 2019.

[140] S. H. Abdulhussain, A. R. Ramli, B. M. Mahmmod, M. I. Saripan, S. A. R. Al-Haddad, and W. A. Jassim, "A new hybrid form of Krawtchouk and Tchebichef polynomials: Design and application," *J. Math. Imag. Vis.*, vol. 61, no. 4, pp. 555–570, May 2019.

[141] W. A. Jassim, R. Mukundan, and P. Raveendran, "New orthogonal polynomials for speech signal and image processing," *IET Signal Process.*, vol. 6, no. 8, pp. 713–723, Oct. 2012.

[142] B. M. Mahmmod, A. R. B. Ramli, S. H. Abdulhussain, S. A. R. Al-Haddad, and W. A. Jassim, "Signal compression and enhancement using a new orthogonal-polynomial-based discrete transform," *IET Signal Process.*, vol. 12, no. 1, pp. 129–142, Feb. 2018.

[143] J. Haddadnia, M. Ahmadi, and K. Raahemifar, "An effective feature extraction method for face recognition," in *Proc. Int. Conf. Image Process.*, vol. 3, Sep. 2003, pp. 3–917.

[144] S. Ghosal and R. Mehrotra, "A moment-based unified approach to image feature detection," *IEEE Trans. Image Process.*, vol. 6, no. 6, pp. 781–793, Jun. 1997.

[145] N. Guo, L. Diao, and Y. Xing, "Projective moment invariants," in *Proc. 4th Int. Conf. Comput. Sci. Appl. Eng.*, Oct. 2020, pp. 1–5.

[146] B. J. Chen, H. Z. Shu, H. Zhang, G. Chen, C. Toumoulin, J. L. Dillenseger, and L. M. Luo, "Quaternion Zernike moments and their invariants for color image analysis and object recognition," *Signal Process.*, vol. 92, no. 2, pp. 308–318, Feb. 2012.

[147] D.-G. Sim, H.-K. Kim, and R.-H. Park, "Invariant texture retrieval using modified Zernike moments," *Image Vis. Comput.*, vol. 22, no. 4, pp. 331–342, Apr. 2004.

[148] D. Rivero-Castillo, H. Pijeira, and P. Assunçao, "Edge detection based on Krawtchouk polynomials," *J. Comput. Appl. Math.*, vol. 284, pp. 244–250, Aug. 2015.

[149] R. Benouini, I. Batioua, K. Zenkouar, S. Najah, and H. Qjidaa, "Efficient 3D object classification by using direct Krawtchouk moment invariants," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 27517–27542, Oct. 2018.

[150] B. Xiao, Y. Zhang, L. Li, W. Li, and G. Wang, "Explicit Krawtchouk moment invariants for invariant image recognition," *J. Electron. Imag.*, vol. 25, no. 2, Mar. 2016, Art. no. 023002.

[151] W. A. Jassim and P. Raveendran, "Face recognition using discrete Tchebichef-Krawtchouk transform," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2012, pp. 120–127.

[152] H. Wu and S. Yan, "Computing invariants of Tchebichef moments for shape based image retrieval," *Neurocomputing*, vol. 215, pp. 110–117, Nov. 2016.

[153] M. El Mallahi, A. Mesbah, H. Karmouni, A. El Affar, A. Tahiri, and H. Qjidaa, "Radial Charlier moment invariants for 2D object/image recognition," in *Proc. 5th Int. Conf. Multimedia Comput. Syst. (ICMCS)*, Sep. 2016, pp. 41–45.

[154] H. Amakdouf, M. El Mallahi, A. Zouhri, A. Tahiri, and H. Qjidaa, "Classification and recognition of 3D image of Charlier moments using a multilayer perceptron architecture," *Procedia Comput. Sci.*, vol. 127, pp. 226–235, Jan. 2018.

[155] K. W. See, K. S. Loke, P. A. Lee, and K. F. Loe, "Image reconstruction using various discrete orthogonal polynomials in comparison with DCT," *Appl. Math. Comput.*, vol. 193, no. 2, pp. 346–359, Nov. 2007.

[156] Z. N. Idan, S. H. Abdulhussain, and S. A. R. Al-Haddad, "A new separable moments based on Tchebichef-Krawtchouk polynomials," *IEEE Access*, vol. 8, pp. 41013–41025, 2020.

[157] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[158] M. Mejdoub and C. Ben Amar, "Classification improvement of local feature vectors over the KNN algorithm," *Multimedia Tools Appl.*, vol. 64, no. 1, pp. 197–218, May 2013.

[159] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31. Berlin, Germany: Springer, 2013.

[160] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[161] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[162] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[163] A. Bousdekis, B. Magoutas, D. Apostolou, and G. Mentzas, "Review, analysis and synthesis of prognostic-based decision support methods for condition based maintenance," *J. Intell. Manuf.*, vol. 29, no. 6, pp. 1303–1316, Aug. 2018.

[164] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[165] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[166] H. Liu and S. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining," *J. Syst. Softw.*, vol. 85, no. 5, pp. 1067–1074, May 2012.

[167] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.

[168] V. Pestov, "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?" *Comput. Math. Appl.*, vol. 65, no. 10, pp. 1427–1437, May 2013.

[169] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing AutoML in educational data mining for prediction tasks," *Appl. Sci.*, vol. 10, no. 1, p. 90, Dec. 2019.

[170] X. H. Cao, I. Stojkovic, and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC Bioinf.*, vol. 17, no. 1, p. 359, Dec. 2016.

[171] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, Clustering, and Data Mining Applications*. Berlin, Germany: Springer, 2004, pp. 639–647.

[172] S. Tsang, B. Kao, K. Y. Yip, W.-S. Ho, and S. D. Lee, "Decision trees for uncertain data," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 64–78, Jan. 2011.

[173] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten, "Using model trees for classification," *Mach. Learn.*, vol. 32, no. 1, pp. 63–76, 1998.

[174] B. Lesser, M. Mücke, and W. N. Gansterer, "Effects of reduced precision on floating-point SVM classification accuracy," *Procedia Comput. Sci.*, vol. 4, pp. 508–517, Jan. 2011.

[175] N. Becker, G. Toedt, P. Lichter, and A. Benner, "Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data," *BMC Bioinf.*, vol. 12, no. 1, p. 138, Dec. 2011.

[176] J. Jin, X. Cai, and X. Lin, "Efficient SVM training using parallel primal-dual interior point method on GPU," in *Proc. Int. Conf. Parallel Distrib. Comput., Appl. Technol.*, Dec. 2013, pp. 12–17.

[177] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Netw.*, vol. 61, pp. 32–48, Jan. 2015.

[178] Z. J. Ding and Y.-Q. Zhang, "Additive noise analysis on microarray data via SVM classification," in *Proc. IEEE Symp. Comput. Intell. Bioinf. Comput. Biol.*, May 2010, pp. 1–7.

[179] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: Fast feature extraction and SVM training," in *Proc. CVPR*, Jun. 2011, pp. 1689–1696.

[180] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 947–954.

[181] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[182] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[183] L. Chen, W. Li, Z. Xu, and L. Tang, "Pedestrian detection based on ISC in infrared images," in *Proc. 3rd Int. Conf. Netw. Distrib. Comput.*, Oct. 2012, pp. 166–169.

[184] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[185] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2009.

[186] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2009.

[187] X. Li, X. Fang, and Q. Lu, "On-road vehicle and pedestrian detection using improved codebook model," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, Jul. 2013, pp. 1–4.

[188] S. Mehralian and M. Palhang, "Pedestrian detection using principal components analysis of gradient distribution," in *Proc. 8th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, Sep. 2013, pp. 58–63.

[189] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 645–659.

[190] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2903–2910.

[191] D. Levi, S. Silberstein, and A. Bar-Hillel, "Fast multiple-part based object detection using KD-ferns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 947–954.

[192] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Aberystwyth, U.K., 2010.

[193] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 24–31.

[194] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 423–436.

[195] J. Kim, J. Lee, C. Lee, E. Park, J. Kim, H. Kim, J. Lee, and H. Jeong, "Optimal feature selection for pedestrian detection based on logistic regression analysis," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 239–242.

[196] B. Li, Y. Li, B. Tian, F. Zhu, G. Xiong, and K. Wang, "Part-based pedestrian detection using grammar model and ABM-HoG features," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, Jul. 2013, pp. 78–83.

[197] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 127–142.

[198] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 241–254.

[199] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[200] S. Yao, T. Wang, W. Shen, S. Pan, Y. Chong, and F. Ding, "Feature selection and pedestrian detection based on sparse representation," *PLoS ONE*, vol. 10, no. 8, Aug. 2015, Art. no. e0134242.

[201] J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, and Y. Burov, "Information resources processing using linguistic analysis of textual content," in *Proc. 9th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Syst., Technol. Appl. (IDAACS)*, vol. 2, Sep. 2017, pp. 573–578.

[202] R. M. Caruso, "Social media content management system and method," U.S. Patent 8 943 054, Jan. 27, 2015.

[203] M. Fanfani, M. Iuliani, F. Bellavia, C. Colombo, and A. Piva, "A vision-based fully automated approach to robust image cropping detection," *Signal Process., Image Commun.*, vol. 80, Feb. 2020, Art. no. 115629.

[204] S. Rothfuss, R. Schmidt, M. Flad, and S. Hohmann, "A concept for human-machine negotiation in advanced driving assistance systems," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3116–3123.

[205] J. Weyer, R. D. Fink, and F. Adelt, "Human–machine cooperation in smart cars. An empirical investigation of the loss-of-control thesis," *Saf. Sci.*, vol. 72, pp. 199–208, Feb. 2015.

[206] J.-L. Yin, B.-H. Chen, K.-H.-R. Lai, and Y. Li, "Automatic dangerous driving intensity analysis for advanced driver assistance systems from multimodal driving signals," *IEEE Sensors J.*, vol. 18, no. 12, pp. 4785–4794, Jun. 2018.

[207] M. M. Rahman, M. F. Lesch, W. J. Horrey, and L. Strawderman, "Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems," *Accident Anal. Prevention*, vol. 108, pp. 361–373, Nov. 2017.

[208] M. Bilal, A. Khan, M. U. K. Khan, and C.-M. Kyung, "A low-complexity pedestrian detection framework for smart video surveillance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2260–2273, Oct. 2017.

[209] A. C. Nazare, Jr., and W. R. Schwartz, "A scalable and flexible framework for smart video surveillance," *Comput. Vis. Image Understand.*, vol. 144, pp. 258–275, Mar. 2016.

[210] P. Birnstill, D. Ren, and J. Beyerer, "A user study on anonymization techniques for smart video surveillance," in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2015, pp. 1–6.

[211] Q. Wang, Q. Ma, C.-H. Luo, H.-Y. Liu, and C.-L. Zhang, "Hybrid histogram of oriented optical flow for abnormal behavior detection in crowd scenes," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 30, no. 2, Feb. 2016, Art. no. 1655007.

[212] X. Zhang, Q. Zhang, S. Hu, C. Guo, and H. Yu, "Energy level-based abnormal crowd behavior detection," *Sensors*, vol. 18, no. 2, p. 423, Feb. 2018.

[213] M. S. Zitouni, H. Bhaskar, J. Dias, and M. E. Al-Mualla, "Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques," *Neurocomputing*, vol. 186, pp. 139–159, Apr. 2016.

[214] Y.-L. Hou and G. K. H. Pang, "People counting and human detection in a challenging situation," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 41, no. 1, pp. 24–33, Jan. 2011.

[215] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1367–1376.

[216] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4099–4108.

[217] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1092–1108, Apr. 2020.

[218] A. Booranawong, N. Jindapetch, and H. Saito, "Adaptive filtering methods for RSSI signals in a device-free human detection and tracking system," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2998–3009, Sep. 2019.

[219] J. Lu, T. Zhang, F. Hu, and Q. Hao, "Preprocessing design in pyroelectric infrared sensor-based human-tracking system: On sensor selection and calibration," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 2, pp. 263–275, Feb. 2017.

[220] W. Zhao, R. Lun, C. Gordon, A.-B.-M. Fofana, D. D. Espy, M. A. Reinthal, B. Ekelman, G. D. Goodman, J. E. Niederriter, and X. Luo, "A human-centered activity tracking system: Toward a healthier workplace," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 3, pp. 343–355, Jun. 2017.

[221] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, Dec. 2015.

[222] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Gener. Comput. Syst.*, vol. 81, pp. 307–313, Apr. 2018.

[223] M. G. Ragab, S. J. Abdulkadir, and N. Aziz, "Random search one dimensional CNN for human activity recognition," in *Proc. Int. Conf. Comput. Intell. (ICCI)*, Oct. 2020, pp. 86–91.

[224] S. Arabi, A. Haghighat, and A. Sharma, "A deep-learning-based computer vision solution for construction vehicle detection," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 35, no. 7, pp. 753–767, Jul. 2020.

[225] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 466–481.

[226] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, early access, Oct. 16, 2019, doi: 10.1109/TAFFC.2018.2874986.

[227] L. L. Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognit.*, vol. 53, pp. 130–147, May 2016.

[228] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D human pose estimation: A review of the literature and analysis of covariates," *Comput. Vis. Image Understand.*, vol. 152, pp. 1–20, Nov. 2016.

[229] M. Hu, Y. Wang, Z. Zhang, and Y. Wang, "Combining spatial and temporal information for gait based gender classification," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3679–3682.

[230] E.-S.-M. El-Alfy and A. G. Binsaadoon, "Silhouette-based gender recognition in smart environments using fuzzy local binary patterns and support vector machines," *Procedia Comput. Sci.*, vol. 109, pp. 164–171, Jan. 2017.

[231] G. Huang and Y. Wang, "Gender classification based on fusion of multiview gait sequences," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2007, pp. 462–471.

[232] J.-H. Yoo, D. Hwang, and M. S. Nixon, "Gender classification in human gait using support vector machine," in *Proc. Int. Conf. Adv. Concepts Intell. Vision Syst.* Berlin, Germany: Springer, 2005, pp. 138–145.

[233] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1259–1267.

[234] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[235] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, vol. 1, no. 2, p. 4.

[236] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3213–3221.

[237] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–151.

[238] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.

[239] P. Bilski, P. Mazurek, and J. Wagner, "Application of k nearest neighbors approach to the fall detection of elderly people using depth-based sensors," in *Proc. IEEE 8th Int. Conf. Intell. Data Acquisition Adv. Comput. Syst., Technol. Appl. (IDAACS)*, vol. 2, Sep. 2015, pp. 733–739.

[240] L. Yang, Y. Ren, and W. Zhang, "3D depth image analysis for indoor fall detection of elderly people," *Digit. Commun. Netw.*, vol. 2, no. 1, pp. 24–34, Feb. 2016.

[241] A. Irtaza, S. M. Adnan, S. Aziz, A. Javed, M. O. Ullah, and M. T. Mahmood, "A framework for fall detection of elderly people by analyzing environmental sounds through acoustic local ternary patterns," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1558–1563.

[242] F. Harrou, N. Zerrouki, Y. Sun, and A. Houacine, "Vision-based fall detection system for improving safety of elderly people," *IEEE Instrum. Meas. Mag.*, vol. 20, no. 6, pp. 49–55, Dec. 2017.

[243] X. Cai, S. Li, X. Liu, and G. Han, "Vision-based fall detection with multi-task hourglass convolutional auto-encoder," *IEEE Access*, vol. 8, pp. 44493–44502, 2020.

[244] N. C. Tay, T. Connie, T. S. Ong, K. O. M. Goh, and P. S. Teh, "A robust abnormal behavior detection method using convolutional neural network," in *Computational Science and Technology*. Singapore: Springer, 2019, pp. 37–47.

[245] K.-E. Ko and K.-B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," *Eng. Appl. Artif. Intell.*, vol. 67, pp. 226–234, Jan. 2018.

[246] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017.

[247] S. Benmakrelouf, C. St-Onge, N. Kara, H. Tout, C. Edstrom, and Y. Lemieux, "Abnormal behavior detection using resource level to service level metrics mapping in virtualized systems," *Future Gener. Comput. Syst.*, vol. 102, pp. 680–700, Jan. 2020.

[248] J. Luo, J. Wang, H. Xu, and H. Lu, "Real-time people counting for indoor scenes," *Signal Process.*, vol. 124, pp. 27–35, Jul. 2016.

[249] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 380–391, Sep. 2009.

[250] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1299–1302.

[251] M. B. Shami, S. Maqbool, H. Sajid, Y. Ayaz, and S.-C.-S. Cheung, "People counting in dense crowd images using sparse head detections," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2627–2636, Sep. 2019.

[252] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 475–491.

[253] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 402–419.

[254] Y. Xu, C. K. Ahn, Y. S. Shmaliy, X. Chen, and Y. Li, "Adaptive robust INS/UWB-integrated human tracking using UFIR filter bank," *Measurement*, vol. 123, pp. 1–7, Jul. 2018.

[255] Z. Fu, P. Feng, F. Angelini, J. Chambers, and S. M. Naqvi, "Particle PHD filter based multiple human tracking using online group-structured dictionary learning," *IEEE Access*, vol. 6, pp. 14764–14778, 2018.

[256] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[257] X. Kong, L. Meng, and H. Tomiyama, "Fall detection for elderly persons using a depth camera," in *Proc. Int. Conf. Adv. Mech. Syst. (ICAMechS)*, Dec. 2017, pp. 269–273.

[258] X. Li, T. Pang, W. Liu, and T. Wang, "Fall detection for elderly person care using convolutional neural networks," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2017, pp. 1–6.

[259] A. Ben Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Syst. Appl.*, vol. 91, pp. 480–491, Jan. 2018.

[260] J. Albusac, D. Vallejo, J. J. Castro-Schez, C. Glez-Morcillo, and L. Jiménez, "Dynamic weighted aggregation for normality analysis in intelligent surveillance systems," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 2008–2022, Mar. 2014.

[261] Z. Chen, Y. Tian, W. Zeng, and T. Huang, "Detecting abnormal behaviors in surveillance videos based on fuzzy clustering and multiple autoencoders," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2015, pp. 1–6.

[262] M. Alvar, A. Torsello, A. Sanchez-Miralles, and J. M. Armingol, "Abnormal behavior detection using dominant sets," *Mach. Vis. Appl.*, vol. 25, no. 5, pp. 1351–1368, Jul. 2014.

[263] W. Ren, G. Li, B. Sun, and K. Huang, "Unsupervised kernel learning for abnormal events detection," *Vis. Comput.*, vol. 31, no. 3, pp. 245–255, Mar. 2015.

[264] G. Shu, G. Fu, P. Li, and H. Geng, "Violent behavior detection based on SVM in the elevator," *Int. J. Secur. Appl.*, vol. 8, no. 5, pp. 31–40, Sep. 2014.

[265] R. M. Tomás, S. A. Tapia, A. F. Caballero, S. Ratté, A. G. Eras, and P. L. González, "Identification of loitering human behaviour in video surveillance environments," in *Proc. Int. Work-Conf. Interplay Between Natural Artif. Comput.* Cham, Switzerland: Springer, 2015, pp. 516–525.

[266] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.

[267] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1160–1168, Jul. 2006.

[268] Wahyono, A. Filonenko, and K.-H. Jo, "Detecting abandoned objects in crowded scenes of surveillance videos using adaptive dual background model," in *Proc. 8th Int. Conf. Hum. Syst. Interact. (HSI)*, Jun. 2015, pp. 224–227.

[269] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[270] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4678–4686.

[271] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5098–5107.

[272] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8042–8051.

[273] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 608–617.

[274] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 1–20.

[275] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[276] Y.-M. Song and M. Jeon, "Online multiple object tracking with the hierarchically adopted GM-PHD filter using motion and appearance," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2016, pp. 1–4.

[277] M. Camplani, A. Paiement, M. Mirmehdi, D. Damen, S. Hannuna, T. Burghardt, and L. Tao, "Multiple human tracking in RGB-depth data: A survey," *IET Comput. Vis.*, vol. 11, no. 4, pp. 265–285, Jun. 2017.

[278] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3127–3136.

[279] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro, "Deep convolutional neural networks for pedestrian detection," *Signal Process., Image Commun.*, vol. 47, pp. 482–489, Sep. 2016.

[280] M. Á. Á. de la Concepción, L. M. S. Morillo, J. A. Á. García, and L. González-Abril, "Mobile activity recognition and fall detection system for elderly people using Ameva algorithm," *Pervas. Mobile Comput.*, vol. 34, pp. 3–13, Jan. 2017.

[281] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang, and S. Li, "RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 511–526, Feb. 2017.

[282] S. Weng, L. Xiang, W. Tang, H. Yang, L. Zheng, H. Lu, and H. Zheng, "A low power and high accuracy MEMS sensor based activity recognition algorithm," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2014, pp. 33–38.

[283] A. Dinh, D. Teng, L. Chen, Y. Shi, C. McCrosky, J. Basran, and V. D. Bello-Hass, "Implementation of a physical activity monitoring system for the elderly people with built-in vital sign and fall detection," in *Proc. 6th Int. Conf. Inf. Technol., New Generat.*, 2009, pp. 1226–1231.

[284] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, pp. 144–152, Jan. 2013.

[285] L. Palafox and H. Hashimoto, "Human action recognition using wavelet signal analysis as an input in 4W1H," in *Proc. 8th IEEE Int. Conf. Ind. Informat.*, Jul. 2010, pp. 679–684.

[286] P. Kelly, A. Healy, K. Moran, and N. E. O'Connor, "A virtual coaching environment for improving golf swing technique," in *Proc. ACM Workshop Surreal Media Virtual Cloning (SMVC)*, 2010, pp. 51–56.

[287] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[288] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1310–1323, Jul. 2011.

[289] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 955–960.

[290] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1365–1372. [Online]. Available: http://www.eecs.berkeley.edu/~lbourdev/poselets

[291] S. A. Rahman, S. Y. Cho, and M. K. H. Leung, "Recognising human actions by analysing negative spaces," *IET Comput. Vis.*, vol. 6, no. 3, pp. 197–213, May 2012.

[292] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: Research and evaluation challenges," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 5, pp. 650–663, Oct. 2014.

[293] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[294] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Syst. Appl.*, vol. 105, pp. 233–261, Sep. 2018.

[295] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, "Robust human activity recognition using smartphone sensors via CT-PCA and online SVM," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3070–3080, Dec. 2017.

[296] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[297] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[298] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 90–97.

[299] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[300] J. Liang, Q. Ye, J. Chen, and J. Jiao, "Evaluation of local feature descriptors and their combination for pedestrian representation," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, 2012, pp. 2496–2499.

[301] S.-T. An, J.-J. Kim, J.-W. Lee, and J.-J. Lee, "Fast human detection using Gaussian particle swarm optimization," in *Proc. 5th IEEE Int. Conf. Digit. Ecosyst. Technol. (IEEE DEST)*, May 2011, pp. 143–146.

[302] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.

[303] J. Xu, D. Vázquez, A. M. López, J. Marin, and D. Ponsa, "Learning a part-based pedestrian detector in a virtual world," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2121–2131, Oct. 2014.

[304] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.

[305] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 618–633.

[306] L. Wang, J. Shi, G. Song, and I.-F. Shen, "Object detection combining recognition and segmentation," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2007, pp. 189–199.

[307] D. Gerónimo, A. D. Sappa, A. López, and D. Ponsa, "Adaptive image sampling and windows classification for on-board pedestrian detection," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2007, pp. 1–10.

[308] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 794–801.

[309] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[310] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[311] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[312] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[313] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[314] C. Li, Y. Huang, H. Li, and X. Zhang, "A weak supervision machine vision detection method based on artificial defect simulation," *Knowl.-Based Syst.*, vol. 208, Nov. 2020, Art. no. 106466.

[315] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "EuroCity persons: A novel benchmark for person detection in traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1844–1861, Aug. 2019.

[316] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 789–792.

[317] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 380–393, Feb. 2020.

[318] R. Voeikov, N. Falaleev, and R. Baikulov, "TTNet: Real-time temporal and spatial video analysis of table tennis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 884–885.

[319] D. Marin, Z. He, P. Vajda, P. Chatterjee, S. Tsai, F. Yang, and Y. Boykov, "Efficient segmentation: Learning downsampling near semantic boundaries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2131–2141.

[320] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma, "A hybrid vehicle detection method based on Viola-Jones and HOG + SVM from UAV images," *Sensors*, vol. 16, no. 8, p. 1325, Aug. 2016.

[321] A. I. Taloba, A. A. Sewisy, and Y. A. Dawood, "Accuracy enhancement scaling factor of Viola-Jones using genetic algorithms," in *Proc. 14th Int. Comput. Eng. Conf. (ICENCO)*, Dec. 2018, pp. 209–212.

[322] D. Zhu and X. Wang, "A method of improving SIFT algorithm matching efficiency," in *Proc. 2nd Int. Congr. Image Signal Process.*, 2009, pp. 1–5.

[323] K. Liao, G. Liu, and Y. Hui, "An improvement to the SIFT descriptor for image representation and matching," *Pattern Recognit. Lett.*, vol. 34, no. 11, pp. 1211–1220, Aug. 2013.

[324] H. Nie, K. Long, J. Ma, D. Yue, and J. Liu, "Using an improved SIFT algorithm and fuzzy closed-loop control strategy for object recognition in cluttered scenes," *PLoS ONE*, vol. 10, no. 2, Feb. 2015, Art. no. e0116323.

[325] A. Li, W. Jiang, W. Yuan, D. Dai, S. Zhang, and Z. Wei, "An improved FAST+SURF fast matching algorithm," *Procedia Comput. Sci.*, vol. 107, pp. 306–312, Jan. 2017.

[326] Z. Yang, D. Shen, and P.-T. Yap, "Image mosaicking using SURF features of line segments," *PLoS ONE*, vol. 12, no. 3, Mar. 2017, Art. no. e0173627.

[327] K. Kesorn and S. Poslad, "An enhanced bag-of-visual word vector space model to represent visual content in athletics images," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 211–222, Feb. 2012.

[328] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time visual concept classification," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 665–681, Nov. 2010.

[329] J. Qin and N. H. C. Yung, "Feature fusion within local region using localized maximum-margin learning for scene categorization," *Pattern Recognit.*, vol. 45, no. 4, pp. 1671–1683, Apr. 2012.

[330] T. Kobayashi, A. Hidaka, and T. Kurita, "Selection of histograms of oriented gradients features for pedestrian detection," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, 2007, pp. 598–607.

[331] Y. Jiang, G. Tong, H. Yin, and N. Xiong, "A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters," *IEEE Access*, vol. 7, pp. 118310–118321, 2019.

[332] J. Fehr and H. Burkhardt, "3D rotation invariant local binary patterns," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[333] M. Correa, G. Hermosilla, R. Verschae, and J. Ruiz-del-Solar, "Human detection and identification by robots using thermal and visual information in domestic environments," *J. Intell. Robotic Syst.*, vol. 66, nos. 1–2, pp. 223–243, Apr. 2012.

[334] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 333–356, 1988.

[335] C. Cadena, A. Dick, and I. D. Reid, "A fast, modular scene understanding system using context-aware object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 4859–4866.

SHAHRIAR SHAKIR SUMIT received the B.Sc. degree in computer science and software engineering from American International University-Bangladesh (AIUB), in 2016. He is currently pursuing the M.Sc. degree (by research) with the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS (UTP), Malaysia.

He has worked as a Software Engineer for two years. Since 2018, he has been a Graduate Assistant with the Department of Computer and Information Sciences, UTP. His research interests include object detection, tracking, missing values imputation, and UI/UX. His current project is on human detection.

DAYANG ROHAYA AWANG RAMBLI received the B.Sc. degree in computer science from the University of Nebraska, USA, the M.Sc. degree in computer science from Western Michigan University, USA, and the Ph.D. degree from Loughborough University, U.K. She is currently an Associate Professor with the Department of Computer and Information Science, Universiti Teknologi Petronas, Malaysia. Her primary research interests include virtual reality and augmented reality for health, education and trainings, focusing on information visualization and presentation, multimodal interactions and interactive systems, and user experience.

SEYEDALI MIRJALILI (Senior Member, IEEE) is currently an Associate Professor with the Centre for Artificial Intelligence Research and Optimization, Torrens University Australia. He is internationally recognized for his advances in swarm intelligence and optimization, including the first set of algorithms from a synthetic intelligence standpoint, a radical departure from how natural systems are typically understood, and a systematic design framework to reliably benchmark, evaluate, and propose computationally cheap robust optimization algorithms. He has published more than 200 publications with more than 27 000 citations and an H-index of 58. His research interests include robust optimization, machine learning, multi-objective optimization, swarm intelligence, evolutionary algorithms, artificial neural networks, and applied optimization.

Prof. Mirjalili is an Associate Editor of several journals, including *Neurocomputing*, *Applied Soft Computing*, *Advances in Engineering Software*, *Applied Intelligence*, *PLOS One*, and IEEE Access. As one of the most cited researchers in artificial intelligence, he is in the list of 2% highly-cited researchers and named as one of the most influential researchers in the world by Web of Science.