Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-10103-4

Human detection techniques for real time surveillance: a comprehensive survey

Mohd. Aquib Ansari[1] · Dushyant Kumar Singh[1]

Received: 10 April 2020 / Revised: 12 September 2020 / Accepted: 19 October 2020 /
Published online: 05 November 2020

Abstract
Real-time detection of humans is an evolutionary research topic. It is an essential and prominent component of various vision based applications. Detection of humans in real-time video sequences is an arduous and challenging task due to various constraints like cluttered environment, occlusion, noise, etc. Many researchers are doing their research in this area and have published the number of researches so far. Determining humans in visual monitoring system is prominent for different types of applications like person detection and identification, fall detection for an elder person, abnormal surveillance, gender classification, crowd analysis, person gait characterization, etc. The main objective of this paper is to provide a comprehensive survey of the various challenges and modern developments seen for human detection methodologies in day vision. This paper consists of an overview of different human detection techniques and their classification based on various underlying factors. The algorithmic technicalities with their applicability to these techniques are deliberated in detail in the manuscript. Different humanitarian imperative factors have also been highlighted for comparative analysis of each human detection methodology. Our survey shows the difference between current research and future requirements.

Keywords Human detection · Feature description · Deep convolutional neural networks · Recent progress · Surveillance · Computer vision

1 Introduction

In recent years, video footages are attracting more attention due to its elaborate applications for humans to detect. The footages that are acquired from the cameras for surveillance are generally accompanied by lower resolution. The vast majority of the scenes caught by a static camera are accompanied with negligible variations in the background. These footages are used to detect, track and analysis of individual behavior for the surveillance. Surveillance is a very crucial topic in computer vision. It can be used in various areas for security purposes like pedestrian detection, driver assistance systems, gender recognition, person

✉ Mohd. Aquib Ansari
mansari@mnnit.ac.in

[1] CSED, MNNIT Allahabad, Prayagraj, UP, India

counting in the dense crowd, etc. These days, the surveillance system is also acting as a crime deterrent that helps to discourage criminals from carrying out illegal activities. Most existing video surveillance frameworks depend on human spectators for identifying particular events in the real-time video sequences. But, there are some impediments in the human ability to observe the concurrent proceedings in the surveillance displays. Therefore, the automatic video surveillance system [7, 22, 28, 50, 113, 127, 152, 168, 189] is needed to detect, track, and examine the individual's behaviors in the videos.

Human detection [5, 39, 63, 69, 84, 101, 108, 119, 120, 141, 159, 180, 186] is an act of localizing all the instances of human being present in an image using computer vision algorithm. With the increasing accentuation on biometrics and surveillance, special attention has been conferred to the task of identifying human beings in images. Surveillance system frameworks are normally intended to screen individuals or persons, both to guarantee security and to detect any unscrupulous act. The performance of monitoring systems frowningly rests on the detection and localization of persons in the videos. The detection of humans is a particularly challenging job due to the articulated and non-flexible nature of the human body. Therefore, the task of human detection turns out to be a more attractive and active research area in the field of computer vision.

In the past, there had a manual method of human detection in surveillance videos. But, that approach had many drawbacks. First, it required someone to continually keep their eye on the video to check for the human presence, and this was not possible by any single person. So, more than one person was to be employed on a shift basis. Second, human negligence was another problem where human beings tend to miss some parts of the video because no one can sit all the time attentively and can't look at the video continuously for several hours. Third, surveillance is a crucial need for any place where security is essential such as Banks, Military basements, Administrative buildings, etc. Even small negligence in the observers may lead to disasters. These drawbacks raised the need for an automated human detection framework [35, 36, 49, 55, 113, 116, 137, 140, 145, 147, 180] where the computer can itself detect human presence and raise an alert if needed. For example, in banks, cameras installed in the front gate will almost always detect the presence of human beings within the bank timing. But beyond these timings, the human presence would trigger an alarm by the system. In another example, some of the military places are very sensitive and restricted to humans. So, if any human is found there, then the alarm would raise by an automated detection mechanism. This will alert the military about any intruder if there is.

Figure 1 shows the general human detection framework. First, the input is taken by the camera module in the form of video sequences. The captured sequences are passed to the object detection module. Object detection is a process of localizing objects within an image or video sequence. Further, these detected objects are classified as human or non-human on the basis of the object classification module. Some of the human detection samples are illustrated in Fig. 2.

It is seen that effectiveness is a crucial need in surveillance situations. So, various human detection algorithms have been proposed depending on accuracy, affordable, and situations. Once human beings are found in the video, they can be tracked and monitored easily. Human

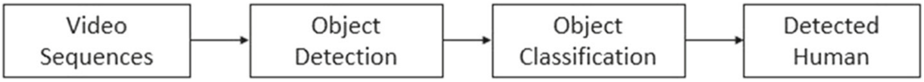


Fig. 1 Human detection framework

Fig. 2 Some samples of human detection framework outcomes

detection is also a basic building block for various applications like human activity recognition [83, 85, 90], pose estimation [29, 45], etc. Here, humans first need to be detected and then processed on the basis of its application.

As we know that humans can identify to another human with the help of some subtle clues. But, human detection algorithms are not up to that level or still far from matching for human detection as humans do. Because the various parameters can be affected due to surroundings at which human is positioned and intrinsic complications linked with the human body. The human body is flexible in nature. So, it yields the number of possible poses. Adjusting the camera position and direction for the appropriate view angle is also a very challenging task [114, 119]. If it is not proper, then it may produce a size variation problem. The additional complexity can be taken place if the human is dressed with different colors and textures. Apart from this, the environment also plays a significant role that may affect the visual appearance of the human. For example, ambient light can increase or decrease the visible spectrum of the human object. The human could be camouflaged by the cluttered background that can often be met with outside scenes. Moreover, whether there is one or more than one person involves in any activity in the occluded or crowded scene where the human body is not fully visible, is indispensable [55, 113, 161, 171].

This article surveys various human detection techniques in brief. The article is organized into five sections. Section 2 illustrates the classification of human detection techniques. In this section, the detailed description of each technique with its importance in real-time has been described in brief. Apart from these, it comprises the summary and comparative analysis part in which the vast comparison among involved human detection techniques has been made. Section 3 contains studies related to standard human detection datasets and different performance measures. Section 4 outlines and compares various researches on human detection. At last, Section 5 provides a conclusion.

2 Classification of human detection techniques

The wide availability of research has been seen in the area of human detection, which has proposed several methods. These methods can be broadly classified based on the nature of the algorithm. The methods can be processing based or learning paradigm based. The processing based techniques perform certain operations (like differencing, curve geometry, etc.) to produce the outcome. The learning paradigm based techniques use machine learning or deep neural architecture to locate humans' presence within an image accurately. The machine learning based techniques allow machines to utilize the experience to advance the tasks. Feature extraction and classification are the two most essential tasks of machine learning. In which, feature extraction is the task to extract important information or attributes (like color, texture, and shape) from an image, and classification is the act of assigning the labels to the objects based on these extracted attributes for decision-making purposes. In contrast, deep learning is a subfield of machine learning that is based on Artificial Neural Networks(ANN) [37, 89, 126, 156, 167]. The structure of an ANN is motivated by the human brain's biological neural network, leading to a process of learning that is significantly proficient than that of standard machine learning paradigms based prototypes.

This article surveys traditional as well as modern developments seen for human detection. Depending on the nature of the work, this article has classified human detection techniques into four types of modules, depicted in Fig. 3. These four modules are described as follows:

- **Face Feature Based:** Face feature based human detectors detect the person's appearance based on their face, which classifies a person's face as a person when they are found within an image.
- **Motion Feature Based:** Motion feature based human detectors trace the pixel movement vectors in the successive frames, and further, these pixels are classified as moving pixels or background pixels.
- **Body Appearance Based:** Body appearance based human detectors use human appearance information such as body shape, curves, and component descriptions to identify the presence of a human in an image.
- **Deep Learning Based:** Deep learning based human detectors employ a single layer or multi-layer deep convolutional architecture to locate humans' presence within an image.

The techniques involved in these modules have comprised of a different principle. For example, in a face feature based module, it is expected that if a face is present within an image, it is classified as human. Face feature based module is divided into three categories: skin color based, face texture based, and face geometry based. A skin color based

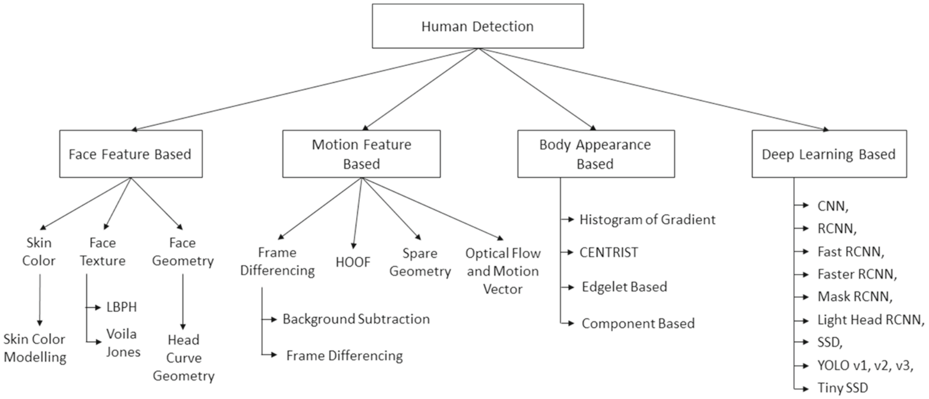


Fig. 3 Classification of human detection techniques

algorithm detects the human's face based on skin pixels. The face texture based algorithm deploys the spatial and textural information to identify the human face. Whereas face geometry deals with different essential patterns like head curves, lines, points, etc. to identify the human face. The face feature based techniques are described in Section 2.1. In the motion based human detection algorithms, all the moving objects are separated from the background frame, and the number of tests are performed to determine whether these moving objects are humans or not. Section 2.2 briefly demonstrates the techniques based on motion features. The body appearance based human detection algorithm seeks the existence of the human body within an image. The human body has likewise be seen as a shape/posture, curves, and components (like two legs, two hands, a central body, and one head). Based on body shape/appearance information, the essential features are mined and used to build a classifier, which identifies a human. The body appearance based techniques are explained in Section 2.3. In contrast, deep learning techniques extract features intrinsically using a single layer or multi-layer deep convolutional neural networks and show some fascinating advancements in the detection process. Deep learning based module includes advanced/modern object detectors (such as R-CNN, Faster R-CNN, YOLO, etc.) to identify the presence of humans [43, 179]. Deep convolutional network based techniques are described in more detail in Section 2.4.

2.1 Face feature based human detection

The number of the algorithm has been proposed in the field of face detection [12, 25, 34, 58, 64, 125, 159, 162]. The face feature based human detection techniques identifies the human on the basis of the human's face. The face feature based human detection includes skin color, face texture, and face geometry based techniques. These techniques are described in brief as follows:

2.1.1 Skin color based face detection

The skin color modeling [4, 5, 19, 27, 42, 47, 61, 79, 102, 154, 183] is an active research area as well as an important preprocessing step for most object detection techniques. The application of the skin detection is ranging from face detection and tracking [33, 136], content based image retrieval (CBIR) [8, 9], gesture recognition and analysis to different Human Machine Interaction (HMI) [146] Domain. The skin color plays a significant role in building an effective face detection system. Finding skin pixels in the visible range of spectrum can be challenging because the skin pixels within an image are sensitive to several factors like camera characteristics, illumination, reflections, ethnicity, individual characteristics, etc. [48, 79, 87, 135, 144, 149]

Detection of skin pixel is an indication of the appearance of human skin within the digital image that transforms the color image into a thresholded binary image. The pixels that have '1' intensity value are labelled as 'Skin pixels' and the remaining pixels are labelled as 'non-skin pixels'. Researchers have offered several ways to deal with this problem, which are as follows:

Skin Color Detection Based on Thresholding For determining the skin color, the thresholding technique is used to cluster the skin color with the help of decision boundaries. The threshold values can be single or multiple, which can be decided through one or more than one color space's components. Based on these predefined or customized thresholding range limits, the image pixel can be defined as a skin pixel or non-skin pixel. Here, color space puts a significant role in deciding the threshold boundaries. Practically, it is seen that skin detection by utilizing multiple color components of more than one color space produces a better result than dealing with single-color space [16, 26, 97, 98, 128, 165].

Hani k. al-mohair [4] used various color spaces for pixel skin detection, e.g., RGB, normalized RGB, HSV, LAB, YCbCr, YDbDr, YIQ, YUV color space. The author finds that YIQ color space is providing better separability between non-skin and skin pixels than other color space. Maheswari S and Reeba Korah [102] used thresholded values for Cb and Cr components of YCbCr color space for detecting the skin color of the human. The range limits used for Cb and Cr color components are [77, 127] and [133, 173] respectively. The skin pixel recognition with a narrow band of values of Cr and Cb color component in YCbCr color space was accounted in [23] for a captured image. In [26, 27], there are two arrangements of static range limits of Cb [133, 173] and Cr [77, 127], which are fixed as a band of the skin pixels. The authors exhibited that the band acquired from Cb and Cr is viable in separating the skin region regardless of the color variation of the skin. But, it is found that this method is limited to deal with different lightning conditions, reflection, occlusion, etc. Khamar Bhasha Shaik et al. [144] have done a comparative study on human skin detection. The author used HSV and Ycbcr color space to detect the skin pixels. They found that the YCbCr color space can work optimally with the rough illumination condition for complex color images. Wang and Yung [172] used RGB and HSV color space with color component's thresholding range of r=[0.36, 0.465], g=[0.28, 0.363], H=[0, 50], S=[0.20, 0.68], V=[0.35, 1.0] to discriminate skin color that classifies skin and non-skin pixels.

Skin Color Detection Based on Classification The classification [14, 73, 78, 99, 155] is a way of categorizing the things in groups according to their likeness. According to the approach of classification, the identity of the skin color can be seen as two-class problems. These classes can be skin pixel and non-skin pixel. With the support of classifiers, the image pixel can be categorized into one of these two classes, which decides whether the pixel belongs to the skin pixel or not. Various classification algorithms are used by different researchers to deal with this problem.

Rehanullah khan et al. [84] presented a comparative study on the classifiers like AdaBoost, BayesNet, J48 tree based, Multilayer perceptron, Naïve Bayes, Random Forest, Radial basis function, Support vector machine and Histogram approach for skin color modeling. The different color spaces (like HIS, RGB, nRGB, CIELAB, YCbCr, IHLS) are also be taken for representing the image. The 8991 different frames are used as the dataset. These frames were taken out from 25 videos, which were downloaded by the internet. These classifiers were applied one by one on the input image with different color spaces. They found that the behavior of different classifiers is varied for different color space. It is also found that the J48 and Random forest classifier for skin color pixel based classification provide excellent results in terms of F-measure.

M.J. Jones et al. [77] constructed an RGB histogram based 3D model. This model was built over 2 billion pixels, which were taken from 18,696 web images. They said that 77% of RGB colors had not been encountered and most of them histograms are empty. However, the ratio of skin pixels to non-skin pixels are $\frac{1}{10}$ only. It is suggested that the color of the skin is more often than the other colors of object. The probability for a pixel, whether it is skin pixel or not, is evaluated as follows.

$$p(\frac{c}{skin}) = \frac{s(c)}{T_s} \tag{1}$$

$$p(\frac{c}{non-skin}) = \frac{n(c)}{T_n} \tag{2}$$

Where $s(c)$ and $n(c)$ represents the total number of pixel of skin histogram and non-skin histogram in the color c-bin respectively. T_s and T_n denote the overall counts in the skin and non-skin bins of the histogram. The Bayes maximum likelihood methodology [44] can be used for building skin classifier.

$$\frac{p(\frac{c}{skin})}{p(\frac{c}{non-skin})} \geq \Theta, \quad for \quad 0 \leq \Theta \leq 1 \tag{3}$$

Where θ (threshold value) between true positive and false negative can be accustomed to the trade-off. According to the author, this proposed method is fast and straightforward because it requires only 2 table look-ups to evaluate the skin pixel probability. G. Gomez et al. [61], Brand and Mason [17], K. Schwerdt, J.L. Crowely [138], S. L. Phung et al. [123] has used **naïve bayes classifier**.

Gaussian classifier [84, 147, 182, 183] is also another option for the classification of skin pixels. It has been seen that many of the researchers have used the Gaussian mixtures model for skin color modeling. One of the benefits of the Gaussian classifier is that it can efficiently generalize with smaller training data. It also requires minimum training requirements. Many varients of Gaussian model are available and can be used according to their applicability.

Yang and Lu [182] used the single Gaussian model for skin color distribution and they found that multivariate Gaussian distribution can model the skin color dissemination for diverse individuals under certain illuminating conditions. Here, the skin color dissemination is modeled as follows.

$$p(c) = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(c-\mu)^T \Sigma^{-1}(c-\mu)] \tag{4}$$

Where $p(c)$ is elliptical Gaussian joint probablity density function, c, μ and σ represents color vector, mean vector and covariance matrix respectively. Yang and Ahuja [183] proved that the single Gaussian distribution could not be effectually modeled for skin color modeling under varying illumination conditions. They used **Gaussian mixture** density function (it is the sum of individual Gaussians), shown in (5), which can describe the distributions of complex shapes.

$$p(c) = \sum_{i=1}^{N} \omega_i \frac{1}{\sqrt{2\pi|\Sigma_i|}} \times \exp[-\frac{1}{2}(c-\mu_i)^T \Sigma^{-1} \sum_{i}^{-1}(c-\mu_i)] \tag{5}$$

Where c, μ_i and Σi are color vector, mean vector and covariance matrix respectively. N and ω are the numbers of Gaussians and weight vector respectively. Jones and Rehg [77] used this model with 16 Gaussians. This Gaussian mixture model has also been used by Reza Hassanpour et al. [70]. J. Y. Lee et al. [91] proposed an alternative variant of GMM called **Elliptical Boundary Model** which has low computation complexity than others.

Multi-layer Perceptron (MLP) classifier [188] can deal with complex non-linear relationships of input and output. It has a capability to generalize any kind of specified data with the help of neurons. The neurons are simple processing elements. The performance of MLP is varied on different factors as like number of node and hidden layers used, rate of learning, etc. The multi-layer perceptron based skin classification is trained to learn the distribution of the complex classes for skin/non-skin pixels. The MLP based skin color modeling has been used by Hani [5], Hongming Zhang et al. [188], Kishor Bhoyar et al. [14]. S.L. Phung et al. [122] applied the MLP for skin color modeling in YCbCr color space.

Brown et al. [18] used the Self-Organizing Map (SOM) classifier for learning the skin/non-skin pixels. SOM is a well-known unsupervised learning based ANN models. They took 500 images for learning. They compared this model with the GMM and they found that SOM produced better results than GMM. The jones and Rehg [77] reported that the performance of SOM might be enhanced by taking a larger number of neurons and more extensive training data.

The other classifiers, which can be used to classify the skin and non-skin pixels, are Maximum Entropy classifier (MaxEnt), Bayesian Network Classifier, SVM classifier, KNN classifier, etc. Sebe et al. [139] used Bayesian Network (BN) for skin-color modeling and classification. Jedynak et al. [76] used Maximum Entropy classifier for classifying the skin and non-skin pixels.

2.1.2 Face texture based human detection

Face texture based human detection extracts the spatial as well as texture information to locate the human's face within an image. The techniques involved in dealing with textural information of face are described as follows:

Voila Jones Technique Voila jones is the first framework to detect the object in real-time competitively, which was proposed by Paul voila and Michael jones [166] in 2001. It is a visual object detection technique based on machine learning. Voila jones algorithm is robust with high detection rates and can easily be deployed in real-time. This method can only detect faces but cannot recognize them. This algorithm is playing an important role in object detection with high detection rate [2].

The algorithm is divided into four essential tasks. The first task uses Haar feature selection technique [33, 118, 136] to extract the essential textural information of the face. The second task introduces the concept of Integral Image (a new image representation method) that allows the easy and quick computation. In the third task, the AdaBoost learning algorithm is used to construct extremely efficient classifiers, in which some critical visual features are selected from a broader set [53]. The crucial fourth task uses a method for combining more complex classifiers in a cascade, which allows removal of background regions and computing the presence of object characteristics from more promising regions. The cascade can be viewed as an object-specific focus-of-attention mechanism, providing statistical guarantees that discarded regions are unlikely to contain the object of interest.

In the first key contribution, the algorithm uses Haar features boxes in varying sizes. There are three kinds of features:

- Two rectangle features aligned in vertical or horizontal or tilted directions, also known as edge features.
- Three rectangle features with two blacks and one white or two whites and one black box, also known as line features.
- Four rectangle features with two black and two white boxes aligned in various directions.

These different types of Haar-like features are also illustrated in Fig. 4. The pixel values are used to sum up the black and white boxes directly in grayscale format and then the difference of the sums is calculated. This difference is compared with the standard threshold value corresponding to the threshold of the specific feature.

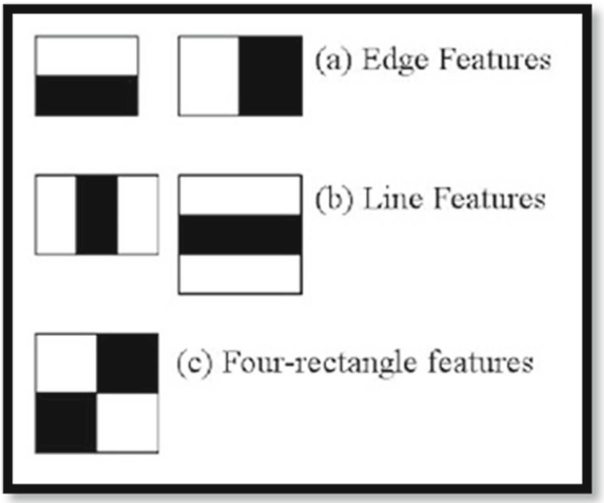


Fig. 4 Haar-like features

With 24 × 24 base resolution, the comprehensive arrangement of the rectangle is somewhat vast, over 180,000. To calculate the value for each Haar feature, the Integral image concept is used to speed up this sum calculation. It evaluates the rectangle features in constant time. This process saves a lot of time. Figure 5 shows the concept of the integral image. At the coordinates P and Q, the integral image comprises the summation of pixels to the left and above the (P, Q), as shown in (6), (7) and (8).

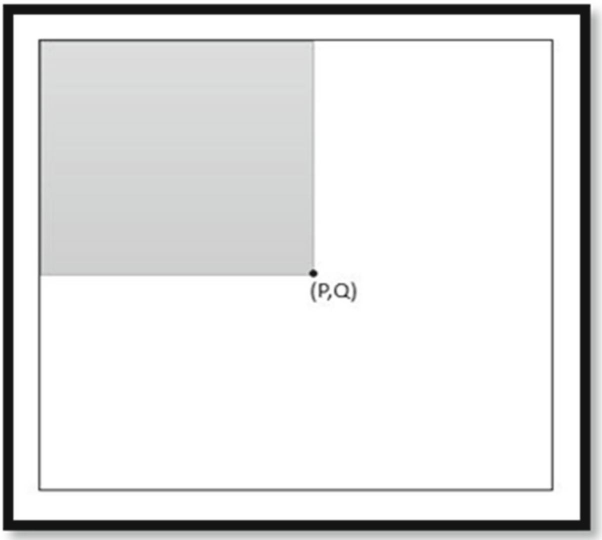


Fig. 5 Integral image

$$II(P,Q) = \sum_{P' \leq P, \quad Q' \leq Q} I(P' + Q') \tag{6}$$

Where II (P, Q) and I (P, Q) represent the integral image and original image, respectively.

$$S(P,Q) = S(P,Q-1) + I(P,Q) \tag{7}$$

$$II(P,Q) = I(P-1,Q) + S(P,Q) \tag{8}$$

Where $S(P,Q)$ is an accumulative row sum. Over the original image, the integral image can be calculated with only a pass for $(P,-1) = 0$ and $II(-1,Q) = 0$.

The third task uses AdaBoost learning methodology [33, 166] to boost up the performance of the classifier. AdaBoost learning methodology provides a kind of formal guarantees. Freund and Schapire [53] has been proved that the training error becomes zero exponentially in large epochs for the robust classifier. It is also to be considered as generalization performance for more number of experiments. The generalization performance is identified with the margins. Therefore, AdaBoost accomplishes the large margins.

In the final task, the single rectangle feature is selected by a weak learning algorithm that helps to classify the negative as well as positive examples efficiently. The cascading is done for better performance and reducing the computation cost. The cascade classifier is the cascade of week classifiers, which helps to form strong classifier. A week classifier contains threshold, parity and features (shown in (9)) specifying the way of inequality.

$$H_k(z) = \begin{cases} 1 & f P_k F_k(z) < P_k \theta_k \\ 0 & otherwise \end{cases} \tag{9}$$

Where $H_k(z)$, F_k, P_k and θ_k represent the week classifier, features, parity and threshold respectively. Figure 6 illustrates the cascade of classifiers.

The promising regions (stages) are identified by the weak features. A number of subwindows are rejected by week classifiers at every stage based on certain criteria. This process not only helps to enhance the next stages of classifier but also increases the detection rate. Now, the strong features that make a strong classifier are cascaded to recognize the characteristics. The voila jones algorithm is much better than another current approach in terms of true positive rate, false positive rate, accuracy, and time complexity.

Local Binary Pattern Histogram (LBPH) The LBP [2, 52] is the simple and most popular texture operation. It labels the pixels by thresholding each pixel's neighborhood of an image

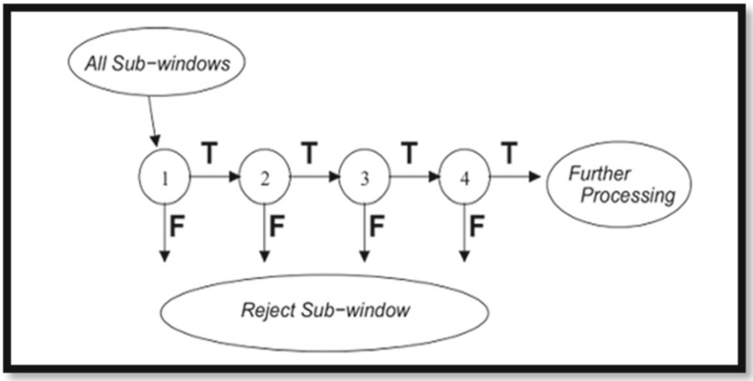


Fig. 6 Cascade of classifiers

and produces the binary numbers. The facial images can be represented with simple features by combining LBP with histograms. One of the objectives of the LBP is to develop the intermediate representation, which can efficiently highlight all the characteristics of the face. To fulfil this objective, the concept of sliding window with neighbourhood relationship is used. The function of LBP operator of 3 × 3 window size is shown in (10).

$$S(x) = \begin{cases} 1 & x \geq 0 \\ 0 & otherwise \end{cases} \tag{10}$$

Figure 7 illustrates the operation of LBP of 3x3 window size (radius=3) with 8 neighbourhood pixels.

The operation of LBP can be extended to a greater number of radius and neighbourhood. This image can be partitioned into the number of grids and for each grid, the histogram can be calculated. Therefore, these calculated histograms can be assumed as a feature vector of LBPH. Then, these feature vectors can be worked as input to one of the classification techniques, which can distinguish the face and non-face.

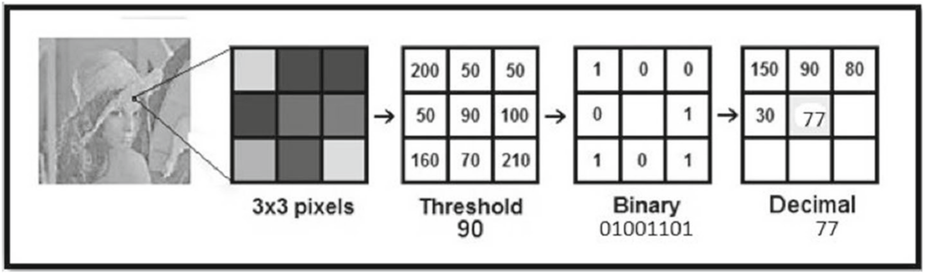
Aftab Ahmed et al. [2] used the linear binary pattern histogram (LBPH) for real-time face detection. They also used the Haar cascade classifier and training recognizer to recognize the face. They found that LBPH is a better way to extract meaningful face features with low time complexity. Firoze and Dev [52] have also used LBPH algorithm to detect the face of the human in real time.

2.1.3 Face geometry based human detection

Human detection based on face geometry uses geometric face parameters such as a straight line, head curve, points, etc. to locate a human face within an image. Face geometry used **Head Curve Geometry** [109] technique to extract the human's face in an image based on their head geometry.

In [175], the Head Curve Geometry technique possesses three steps. These are **pre-processing**, **point searching** and **curve drawing**. In the **pre-processing** step, the thermal image is acquired by the thermal camera at the resolution of 320×240. The reason behind selecting the thermal imaging system is its underlying nature of being resistance to harsh environments. It extracts the red component from the captured image and converts this red component of the image into a binary image according to (11).

$$H_k(z) = \begin{cases} 1 & image_{red}(i, j) > thresold \\ 0 & otherwise \end{cases} \tag{11}$$



200	50	50
50	90	100
160	70	210

1	0	0
0		1
1	0	1

150	90	80
30	77	

Fig. 7 LBP process

Where i and j are the row and column matrix. After generating the binary image, the morphology closing operation and filling holes operation are performed. The **point searching** step is used to find out the various points from the pre-processed image. These points are head point, right-most point, left-most point, right mean, left mean, right neck point, left neck point and i-distance. The i-distance is calculated between the top-most point and extreme points. The **curve drawing** step is used to draw the curve on the basis of points that were found in the previous step according to the algorithm. This algorithm is capable of achieving accuracy up to 91.4 percent. But it needs to be improved to deal with multiple faces. Wong et al. [174] also used the Head Curve for real-time surveillance.

The comparative analysis of Face feature based human detection techniques are illustrated in Table 1, including the basic concepts behind method, usability, pros, cons, and performance. The performance of skin color based techniques may vary when working with different color spaces. Several environmental conditions (such as lousy lightning, reflection, dark background, etc.) may also affect their performance. If these techniques are used with some classifiers, their performance can be increased. On analysing different researches, it is found that Voila jones framework is a popular face detector, which is 15 times faster than other traditional object detectors. LBPH is not much faster than Voila Jones, but it can recognize facial texture information more efficiently. The head curve geometry method can deal with thermal images but cannot detect multiple humans.

2.2 Motion feature based human detection

The performance of face detection based human detection techniques is limited because if a human face is not accurately detected, the framework will not classify it as human while it is a human. The motion based human detection algorithm overcomes this problem by classifying humans based on pixel movement or motion. It includes different techniques such as Frame differencing, Optical flow and motion vector, HOOF, and space geometry, which are described in brief as follows:

2.2.1 Frame differencing based human detection

In video sequencing analysis, the moving object detection is a difficult part. Many of the algorithms have been proposed for analysing the moving object/individual in the video sequences. The pixel based differencing methodologies have drawn the wide separate attention of the researchers for detecting the moving object with the help of static camera in real-time. These pixel based differencing methodologies are described as follows:

Background Subtraction Technique The background subtraction algorithm [62, 82, 124] is one of the simplest algorithm of frame differencing. It is a prevalent methodology to find the moving objects in the video sequences. According to the algorithm, it first evaluates the subtraction operation between the current frame (fr_i) and the background frame (fr_{back}). Then it compares the evaluated pixel values with a predefined threshold value (T). The pixels whose values exceed the predefined threshold value are classified as foreground pixels or moving pixels. The basic operation of the background subtraction method is shown in (12).

$$|fr_i - fr_{back}| > T \tag{12}$$

Where T is the threshold value that helps to distinguish the foreground pixel from background pixel. After the background subtraction process, the resulting image may accumulate with noise. Therefore, it also requires a post-processing step to overcome noise.

Table 1 Comparative analysis of face feature based human detection techniques

Technique	Concept	Applicability	Advantage	Limitation	Performance
Skin Color Modeling [79]	It extracts the face based on skin color.	Skin color modeling is used where storage requirements and computational cost is important. It can be used as a pre-processing step.	It is fast and straightforward. The classifiers can improve the performance of skin detection	Sensitive to illumination, camera characteristics, cluttered background, ethnicity, etc.	Faster computation; Easy to implement; Requires less storage
Linear Binary Pattern Histogram (LBPH) [2]	First, linear binary pattern (LBP) evaluates the features, and then the histogram is formed on these features that help to build to classifier for human face detection.	It depends on the camera position where human faces can be fully covered like above the door.	Provides effective frontal face features in low computation time.	Effective only for frontal face images. Limited to deal with head pose change, occlusion, illustration, larger deformation of the face under varied expression.	Provide more accurate results but not as fast as skin color based techniques.
Voila Jones [166]	It mainly uses haar feature selection, Adaboost training and Cascade classifier.	It is suitable when the camera is positioned in places where human faces can be directly seen, like above a door, etc.	Faster and accurate; Scale and location invariant.	Effective only for frontal face images; Sensitive to deal with varying lighting conditions, cluttered backgrounds and harsh environments.	Face detection proceeds at 15 fps.
Head Curve Geometry [175]	Detects humans by performing the following steps: Pre-processing, point searching and curve drawing.	It is best when there is exactly one human object present within the image.	Fast and effective algorithm; It can deal with various lighting conditions.	It cannot detect multiple humans in the same frame.	Requires low computational cost and can work well with thermal imaging systems.

Figure 8 shows the overall process of background subtraction methodology for human detection. It should be remembered that there should not have any moving object while estimating the background frame at time t. This background image should be updated regularly in such a way that varying luminance and geometry conditions could be adapted properly. The result may be badly affected in case of the poor background image. The many other barriers are possible for effecting the results of background subtraction method like illumination variation, reflection, moving objects, noise, etc. The subtraction technique should be robust in such a way that it can handle all of these limitations [62]. Various extended versions of the background subtraction algorithm are as follows:

Deepjoy and Sarat [38] have studied the 3 types of background subtraction technique for real-time tracking of the moving object. They also performed the comparative analysis among these techniques on the basis of accuracy, memory requirements and speed.

T Horprasert, et al. [74] proposed an extended methodology based on background subtraction technique. This methodology used RGB color image components. This algorithm separates the brightness from the chromaticity component on the basis of the proposed computational color model. Let i is the pixel in an image;

$$\alpha_i = [\alpha_R(i), \alpha_G(i), \alpha_B(i)] \tag{13}$$

$$\beta_i = [\beta_R(i), \beta_G(i), \beta_B(i)] \tag{14}$$

Where, α_i is the i^{th} pixel in the background frame. β_i is i^{th} pixel in the current frame. This β_i is subtracted from the α_i. The main aim is to find out the distortion of current frame (β_i) from the background frame (α_i) in the form of brightness and chromaticity distortion that is shown in (15) and (16) respectively.

$$\phi(\gamma_i) = (\beta_i - \gamma_i.\alpha_i)^2 \tag{15}$$

$$\delta_i = \| (\beta_i - \gamma_i.\alpha_i) \| \tag{16}$$

Where the pixel brightness strength is represented by γ_i.

Stauffer et al. [153] suggested a new approach for real-time tracking based on the adaptive background fusion model. According to this approach, the pixel process needs to be known. The pixel process of pixel I is denoted as $\{I_1, \ldots I_j, \ldots I_t\}$. Where the range of

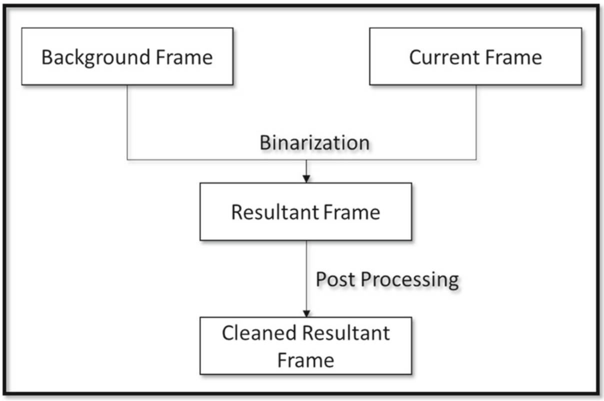


Fig. 8 Background subtraction process

j is more than and equal to 1 and lesser than and equal to t. The probability function for observing the pixel I at the j^{th} frame is depicted in (17) and (18).

$$P(I_j) = \sum_{j=1}^{k} w_{j,t} \times \gamma(M_t, \mu_{j,t}, \Sigma_{j,t}) \tag{17}$$

$$\gamma(M_t, \mu_{j,t}, \sum{}_{j,t}) = \frac{1}{\sqrt{|\Sigma|}\sqrt[n]{2\pi}}.e^{-\frac{1}{2}(M_t-\mu_t)^{\gamma}\sum^{-1}(M_t-\mu_t)} \tag{18}$$

Where $\mu_{j,t}$ and $\sum_{j,t}$ represented as mean and covariance of j^{th} Gaussian mixture at the time instant of t. k represents the number of Gaussians that are present in the mixture. The weight vector can be updated, as shown in (19) to (22).

$$w_{i,j} = (1-\beta) * w_{j,t-1} + \beta * M_{j,t} \tag{19}$$

$$\mu_t = (1-\rho) * \mu_{t-1} + \rho * X_t \tag{20}$$

$$\sigma_t^2 = (1-\rho) * \sigma_{t-1}^2 + \rho * (X_t - \mu_t)^T * (X_t - \mu_t) \tag{21}$$

$$\rho = \beta * \gamma(X_t, \mu_{t-1}, \Sigma_{t-1}).\beta \tag{22}$$

Where β is represented as learning parameter.

This adaptive method overcomes various limitations where traditional background subtraction becomes unstable. But, it is limited to work with varying illumination and object overlapping problems.

Frame Differencing Technique The frame differencing methodology [62, 65, 82, 153] is also used to detect moving objects with the help of the captured frame from the stationary camera. According to the algorithm, the current frame (fr_i) is subtracted from the previous frame (fr_{i-1}). If the pixel value (after calculating difference) is more than threshold (T) value, this pixel will be contemplated as foreground pixel (shown in (23).

$$|fr_i - fr_{i-1}| > T \tag{23}$$

This method is widely adaptive and has the lowest computation cost. One of the limitations of this algorithm is that the object should move continuously. If the object stops its movement, it turns into the background. Additionally, this approach is more prone to deal with noise and movements in the background. Jiajia Guo et al. [65] proposed an improved frame difference, which is shown in (24).

$$D(x, y, \Delta t) = |f(x, y, t) - f(x, y, t-1)|(+)|f(x, y, t+1) - f(x, y, t-1)| \tag{24}$$

Where $D(x, y, \Delta t)$ denotes the differenced image. $f(x, y, t)$, $f(x, y, t-1)$ and $f(x, y, t+1)$ are the 3 adjacent sequences of the video sequence. This enhanced three-frame difference method can detect moving objects more accurately than the traditional frame differencing method.

2.2.2 Optical flow and motion vector

Optical flow [67, 75, 80, 82, 141, 151, 157, 158] is a very popular algorithm of moving object detection in real-time video sequences. It evaluates the motion between the two frames for each pixel at different time intervals. The algorithm aims to separate the foreground/moving objects from the background image and produce optical vectors for the moving object. However, this method is also useful in object tracking, which is gradually used to find object position in all frames. In object tracking, the object of interest is marked in the first frame and then they are tracked in each successive frame using optical flow. The sudden changes in the background can be compensated by taking the mean of every frame in the grey-scale format. At different time intervals t, the motion between two frames is evaluated for each image pixels by the optical flow.

Figure 9 illustrates the point's movement with their orientations. Suppose the object is moved to δx and δy in the x and y direction, respectively, in δt time. It is assumed that the brightness information of the background is constant. At time t, Let $I(x, y, t)$ is the brightness or intensity information at coordinate (x, y) within the image. In the case of pattern movement, the brightness information in the pattern for a particular point is constant. From this statement, it is concluded that $\frac{\delta i}{\delta t} = 0$. The (25) shows the simplified teller series expression, which denotes the two dimensions dynamic intensity or brightness function of $I(x, y, t)$.

$$I_x.v_x + I_y.v_y = -I_t \tag{25}$$

From (25), it can be expressed as:

$$\bigtriangledown I.\overline{v} = -I_t \tag{26}$$

Where $\overline{v}$ and $\bigtriangledown I$ represent the optical flow vector and spatial gradient of intensity respectively.

The moving objects are tracked by the motion vector estimation technique that estimates the object's position in the consecutive frames. To estimate the object's movement, it produces a 2D vector by comparing the consecutive frames. It can eliminate various limitations faced by traditional optical flow. The motion vector estimation algorithm works on the macros blocks of the reference frame [75, 141]. The interested area block of the target frame is compared with the area of reference image (shown in (27)) that produces Mean Absolute Difference (MAD).

$$MAD = \frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} Tar(x+k+i, y+l+j) - Ref(x+k, y+l) \tag{27}$$

Where N, i and j are macros size, horizontal and vertical movement respectively. Indices for pixels within the block are represented by vector k and l. $Tar(x+k+i, y+l+j)$ and $Ref(x+k, y+l)$ denote the pixels present in macro block of target and reference frame respectively.

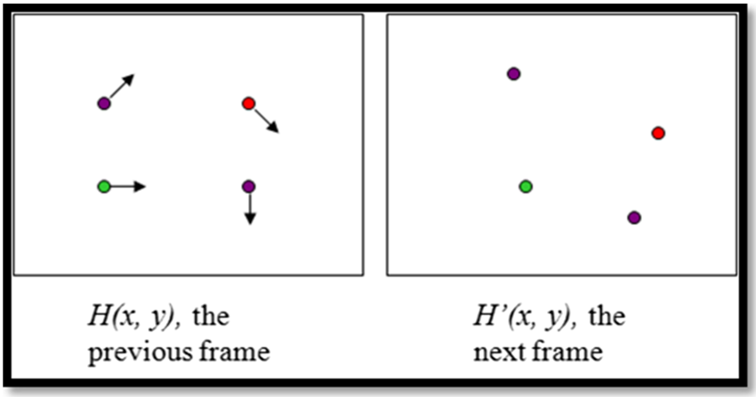


Fig. 9 Optical flow points in two consecutive frames

Kiran Kale, Sushant Pawar, Pravin Dhulakar [80] used optical flow algorithm in conjunction with a motion vector estimation technique for moving object detection. They found that optical flow evaluates the vital information about an object's movement without the computation of the quantitative parameters. Whereas motion vector estimation estimates the position of the object in a continuous frame. Optical flow with motion vector estimation technique not only increases detection accuracy, but also efficiently tackles the blurred and cluttered backgrounds. It can be employed in many different areas like 3D reconstruction, activity recognition, traffic surveillance, medical imagine, human motion analysis, etc.

2.2.3 Histogram of oriented optical flow (HOOF)

In video analysis, it has been found that human tracking, especially in a large event, is a very arduous task. The tracking in an obstructed high-density atmosphere does not put a great significance. The tracked information is valuable in manipulating crowded behavioural standards like paths, speed and density. To deal with this problem, Histogram of oriented optical flow (HOOF) [41, 75, 80, 157] can be deployed with the intention of efficient object tracking. It combines the benefits of motion and shape features. HOOF outputs a histogram vector derived from optical flow. The histogram vector consists of a set of optical flow direction values distributed over several bins such as 8 or 12 or more bins. Further, these vectors along with a classifier determine the moving objects in the video sequence.

Dollar et al. [41] suggested an advanced framework of human detection by incorporating visual features with motion information. In various computer vision application, the motion feature is widely used. Fu-Chun Hsu, Jayavardhana Gubbi and Marimuthu Palaniswami [75] proposed an algorithm to detect people's head and shoulder by combining motion and visual features in low-resolution videos. It can also efficiently deal with occlusion and chaotic environment. They offered a better version of histogram of oriented optical flow (HOOF) named integral histogram of oriented optical flow (Integral-HOOF). Garcia-Martin [57] used the motion MISM and visual MISM features for human detection in crowd. Senst et al. [141] utilized the Gaussian mixture motion model on motion vectors for object detection. This algorithm segments and evaluates the foreground based on optical flow and histogram based on dense optical flow. The authors showed that the combination of motion HOOF and visual HOG together is found to have a better performance of the system in different conditions [36, 170].

2.2.4 Spare geometry based technique

The traditional image processing frameworks use geometrical regularities for the image's feature representation based on edge information. However, these features cannot deal with complex situations. Therefore, G. Peyre et al. [121] presented the geometric flow framework for moving object detection. It described the geometric regularities of the images with Bandelet transform (first generation) in 2000. But, there were many complex operations involved that causes a higher computational cost for evaluating the result. In addition, it provides boundary artifacts because of the non-orthogonal transformation. These problems are solved in 2004 by the second generation of Bandelet transform. In which, the image has been decomposed with the help of geometric flow and multi-scale analysis.

Hong Han, Minglei Tong [68] have proposed a human detection algorithm by detecting the motion of human beings using region segmentation and classification through machine learning. This proposed methodology is based on optical flow and Bandelet transform. This algorithm first evaluates the region segmentation on the basis of optical flow and then

Optical flow based Bandelet transform is used for feature extraction and classification purposes. They found that the bandlet transform evaluates the optimal parameters and reduces the non-zero coefficients in the feature extraction and image compression process.

A comparative analysis of the motion feature based human detection framework is depicted in Table 2. The motion feature based techniques can deal with human detection process very efficiently. They can detect humans when they are in motion. Many applications (such as surveillance, pedestrian detection, human activity recognition, etc.) have used the motion based techniques. The performance of motion based techniques can be affected by different unbiased conditions, such as poor lighting, occlusion, background objects, and noise.

2.3 Body appearance based human detection

Human body appearance based attributes/features are mainly used in image analysis for object recognition. These attributes are invariant to scaling, rotation, and translation. The body appearance based detectors explore the presence of the human on the basis of the appearance/shape pattern of human body. The HOG, CENTRIST, Edgelet, and Components based algorithms are used to mine the appearance information from the image. A detailed description of these techniques is as follows:

Histogram of Oriented Gradient (HOG) The Histogram of Gradient (HOG) [35, 36, 64, 82], a well-known descriptor, is used to detect objects in image processing. It can be easily deployed with the human detection framework by evaluating the human contours within the video sequences. The main objective of HOG is to evaluate the incidence of gradient/slop orientation in local areas of an image. It not only provides the appearance of local objects, but also describes the shape of the objects in an image with the help of edge directions. These edge directions are also be called intensity gradients distribution. The algorithm of HOG is divided in to three modules: a) Gradient Computation, b) histogram Generation, c) Normalization. In the gradient computation module, the image is divided into the number of the blocks and further each block is divided into the number of cells. These cells are small connected regions. The direction of edges is evaluated for all the pixels of each connected region. The horizontal gradient (G_i) and vertical gradient (G_j) are evaluated for each pixel at location (i, j). For each pixel, magnitude and orientation are calculated as given in (28) and (29).

$$Mag_{(i,j)} = \sqrt[2]{G_i^2 + G_j^2} \tag{28}$$

$$\theta_{(i,j)} = tan^{-1}(\frac{G_i}{G_j}) \tag{29}$$

In the next module, the histogram is calculated for each cell on the basis of gradient directions and gradient magnitudes. These gradients are sensitive to different light conditions. To overcome this problem, the normalization task is used so that gradients are not affected by light changes. Normalization normalizes cell's histogram values in a fixed range. The normalized cell histograms are grouped together that denotes to block histograms. Finally, a descriptor is formed by collecting all these block histograms. This descriptor can deal with varying lighting conditions, cluttered backgrounds, and partial occlusion problem to some extent.

L. Greche and N. Es-Sbai [64] used HOG descriptor for facial expression recognition. Whereas Dalal N. et al. [36], Seemanthini K, and Manjunath [140], X. Wang [170] and Yanwei Panga et al. [117] used HOG for human detection.

Table 2 Comparative analysis of motion feature based human detection techniques

Technique	Concept	Applicability	Advantage	Limitation	Performance
Background Subtraction [56]	It subtracts the current or recent frame from the fixed background frame.	It works with static low-resolution cameras having a constant background. Need to ensure that the only intrusion can be made by humans, not by other moving objects.	It is a direct pixel based subtraction technique, which is fast and easy to implement.	Moving objects in the background can severely affect results; Difficult to work with different weathers, illuminations, reflection, moving objects in the background, etc.	Faster and computational efficient
Frame Differencing [65]	It subtracts the current or recent frame from the previous frame.	Applicability is similar to Background Subtraction Technique.	It can resist light interference up to some extent. It is better than background subtraction because it uses consecutive frame of the video sequence.	If an object stops to move, then this algorithm turns it into the background. This approach is more prone to noise, illumination and movement in the background.	Faster and computational efficient
Optical Flow and Motion Vector [67]	It uses optical flow with motion vector estimation for object detection and tracking in the video frame sequences.	It can be used to track more distant objects from the camera with a wide speed variation accurately.	It provides an estimation of object position from consecutive frames, which increases the accuracy and provides robust results irrespective of noise image.	Limited to work with large motion.	Not much faster than Frame Differencing based techniques.
Histogram of Oriented Optical Flow (HOOF) [75]	Detects human by motion extraction, segmentation and histogram building followed by the classification	Cameras positioned in low ceiling corridors may result in severely occluded videos, such videos can work well with HOOF.	It is scale-invariant and does not depend entirely on motion directions; It combines the advantages of motion and shape features.	Optical flow calculations are susceptible to noise and illumination changes.	It takes time to produce a histogram on optical flow.
Spare Geometry [121]	Detects human by region segmentation and classification.	It works for a wide range of camera positioning and learning algorithm needs to be modified accordingly.	It detects human beings out of all moving objects in a video at low time cost and can be used in real-time.	Differet lightning conditions, reflections, occlusion, and noise may produce some false positive instances in an image.	Low Computational cost

Census transform histogram (CENTRIST) The Census Transform (CT) was suggested by Zabih and Woodfill [185]. It is based on the principle of structure kernel. CT is a local non-parametric transform. It tends to be supposed as a comparatively systematic set of intensity between the central element and its neighbourhood, in which the neighbourhood size is not constrained. Figure 10 shows the mechanism of the census transform in 3×3 neighbourhood pixels. Census transform is generally used for local feature extraction from the image locally [39, 109, 133]. For a particular pixel at location (x_c, y_c), the Census Transform is evaluated as given in (30) and (31).

$$CT = \sum_{k=0}^{k-1} f(g_c - g_k) \tag{30}$$

$$f(i, j) = \begin{cases} 0 & I(i, j) > consideration \quad pixel \quad neighbouring \quad eight \quad pixels \\ 1 & otherwise \end{cases} \tag{31}$$

Where I is an image with spatial coordinates i and j. There are 2 limitations in the Census transform.

- It yields irregular histogram distribution.
- It is very difficult to deal with high dimension sparse database.

Census Histogram (CH) [30, 109, 184] illustrates the structure kernel distribution and it is formed from the features of the Census. It is robust and computationally efficient. As is known, the histogram cannot deal with spatial information efficiently. So, to fulfil this characteristic, it is evaluated for small regions of the image. The histogram for region $'m'$ can be evaluated as given in (32).

$$H_{m,i} = \sum_{(x,y)\epsilon R_m} I_i(C_l(x, y)) \tag{32}$$

Where $C_l(x, y)$ represents the labelled Census image, i denotes histogram bins and $I_i(C_l(x, y))$ denotes the indication function to i^{th} bin. At last, these evaluated information descriptors are used to build a classifier for human detection.

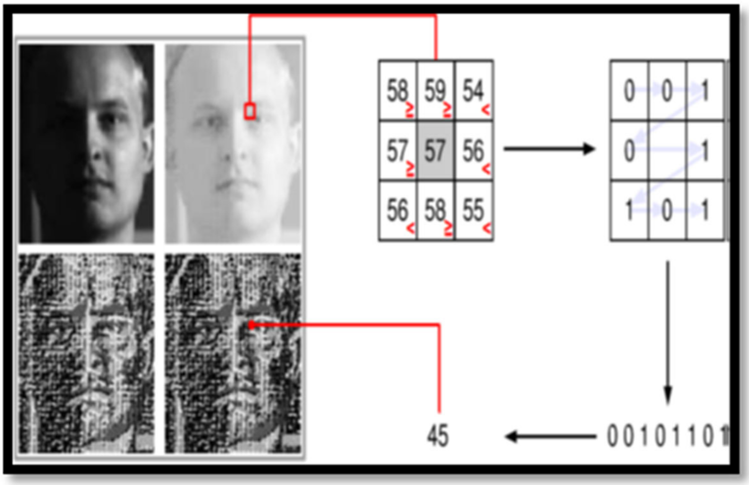


Fig. 10 Census transform mechanism [30]

Census Transform Histogram (CENTRIST) [30, 184] is the histogram of Census Transform (CT). It can be used in various applications such as face detection, human detection, pedestrian detection, etc. The CENTRIST feature descriptor reduces the limitations of the traditional Census transform and provides better information for human detection.

Irfan Riaz, Jingchun Piao and Hyunchul Shin [133] proposed an efficient scheme for human detection by using the thermal images. Here, the census transformation histogram is used for feature extraction and the support vector machine is used for classification purpose. According to Wu, J. et al. [178], The CENTRIST visual descriptor is a very efficient and fast technique for human detection using human contours. It can not only deal with large-scale contours but can also code sign information efficiently. It marely works on 20 fps of 640x480 resolution for human detection. Mu, Y. et al. [109] used the census transform on the IR image directly. They found that CENTRIST descriptor can extract the contours more efficiently than HOG.

Edgelet Based The Edgelet feature [15, 54, 177] is robust scheme for detecting humans in the real-time video sequences. It is invariant to shadows and reflections. The Edgelet is a small piece of curve or line. In Edgelet (E), the normal vectors and positions of the points are denoted by $\left\{n_j^E\right\}_{j=1}^{k}$ and $\left\{u_j^E\right\}_{j=1}^{k}$. Where E, k and I represent edgelet, edgelet length and image respectively. For specified image (I), the intensity of edge as well as normal vector are denoted by $M^l(p)$ and $n^l(p)$ respectively at location p of I. The information of intensity and shape of edges are captured by edgelet affinity function. At w position, the affinity function is shown in (33), which is evaluated between E and I.

$$f(E; I, w) = \frac{1}{k}\sum_{j=1}^{k} M^I(u_j + w)|\left\langle n^I(u_j + w), n_j^E \right\rangle| \tag{33}$$

Where w and u_j are the offset and the coordinate frame of the sub-window within an image respectively. The Sobel kernel (size. 3×3) convolution with grey image gives $M^l(p)$ and $n^l(p)$. BO and Ram [177] used edgelet features to represent the different body parts of the human efficiently like head, shoulders, legs and middle part of the body. Edgelet based boosted classifier [54] is a low computational cost classifier, which can be used with edgelet features to identify the human presence. According to Bhuvaneswari and Rauf [15], the Edgelet features are appropriate for the discovery of humans because they are invariant to clothing differences.

Component based The Component based human detection technique [24, 73, 107] is a good approach for locating the human within the video sequences. The algorithm's main objective is to detect full humans by detecting different body parts like head, arms, and leg. These components need to be present in the appropriate geometric configuration. Therefore, some geometric constraints are used over there, which helps the system to locate the person appropriately. This properly makes this method a visual-immutable.

Chakraborty et al. [24] used the component based technique for human detection within the video sequences. This proposed methodology is applied in the chaotic scenes to detect and locate the human where they are carrying out some actions like running, walking, etc. The algorithm is designed to detect 3 kinds of components such as arms, legs and head for human detection. Here, histogram of gradients descriptor is used to extract the essential gradient's features from the image. The standard deviation based features selection routine is used for these gradient's features where the threshold value is more significant than the gradient's standard deviation. The most substantial candidate element is evaluated based on the component score. The element, which having higher component score, is decided as the most influential component. This system used 4 arm detector, 4 head detector and 1 leg detector. The view angle for the head detector is ranging from 450 to 1350, 1350 to 2250, 2250 to 3150 and 3150 to 4550. The data vectors are generated by feature extraction and feature selection method and these are used by SVM classifier for classification. The optimal hyperplan is evaluated as

$$f(x) = sgn(g(x)) \tag{34}$$

$$g(x) = (\sum_{i=1}^{1*} y_i \alpha_i K(x, x_i^*) + b) \tag{35}$$

Where K represents the kernel functions, x_i is the data points and y_i is a class label with range $\varepsilon \{-1, 1\}$.

A comparative analysis of the body appearance based human detection framework is depicted in Table 3. On analysis, it is found that HOG is an efficient descriptor for object detection but produces large feature lengths, which lead to high computational costs. CENTRIST extracts the attributes locally from the image, and it is much faster than the HOG descriptor. The edgelet descriptor is invariant to reflection, shadows, and clothing differences, but it cannot deal with harsh environments efficiently. Component based human detection technique requires vast computational constraints because it detects each human body component separately and uses geographical configuration constraints to identify the full human body. It can deal with inter-object occlusion as well as self-occlusion more efficiently compared to others.

Finally, on analyzing all feature based human detectors, it has been observed that most human detection techniques produce feature vectors, which require further classification tasks for decision-making purposes. So far, researchers have used a wide variety of classifications to classify objects as human or non-human, which are as follows:

Nemanja and Ljubomir [120] deployed the Naïve Bayes classifier for human tracking. Naïve Bayes classification is based on Bayes theorem, which identifies the probability of one event by looking at the probability of another event. Naïve Bayes follows the principle of predictor's liberation, which means that the features present in a particular class don't depend on the features of another class. It is an extremely sophisticated classification algorithm that easies to build and can handle large datasets efficiently. Yang, Yiqun and Yan [181] used the Naïve Bayes classifier to detect multiple humans for a Binary Pyroelectric Infrared Sensor Tracking System. Luo, Xiaomu et al. [100] also used Naïve Bayes classifier for human activity recognition.

The support vector machine [6, 46, 73] is a standard tool for the object classification task. It is a discriminative classifier stately define by separating hyperplane, which means that the SVM classifier yields a hyperplane to classify input data. SVM uses the kernel function for determining the decision boundaries. The kernel function can be different types like linear, non-linear, polynomial, radial basis function (RBF) and sigmoid. RBF is mostly used kernel function, which is very useful for dealing with higher dimensions. Fu-Chun Hsu et al. [75], Han and Tong [68], Bhaskar Chakraborty et al. [24], Seemanthini and Manjunath [140], Felip and Helio [39] deployed SVM classifier for human detection framework. Suman Kumar Choudhury et al. [31] used another variant of SVM called HIKSVM to detect pedestrians.

Chien-Liang et al. [94] employed K-nearest neighbour classifier (KNN) for human fall detection system. KNN acts as a standard classifier with the assumption that related things

Table 3 Comparative analyses of body appearance based human detection techniques

Technique	Concept	Applicability	Advantage	Limitation	Performance
Histogram of Oriented Gradient (HOG) [36]	Detects humans by applying gradient generation, histogram formation, and normalization process to evaluate features, and then these features are used to build a classifier.	It works well with occluded, as well as clearly defined images.	It works on local cells and it is invariant to geometric and photometric transformations.	The feature length of the HOG is very large, which causes high computational cost.	Requires larger time to calculate features than others
Census Transform Histogram (CENTRIST) [178]	Detects humans by evaluating histogram features on Census transform, and these features are used to build a classifier.	It can be used with normal as well as thermal cameras.	Easier to implement, faster and require fewer parameters to tune. Significant detection performance on various standard datasets.	Sometimes, it yields irregular histogram distribution. It is very tough to deal with high dimension sparse database.	Faster and more accurate than HOG; It can process ~20 frames in a second.
Edgelet Based [15]	Detects humans by evaluating Edgelet features and these features are used to build a classifier.	It is applicable to a wide range of camera variation as well as thermal cameras.	It is invariant to reflection, shadows and clothing differences. It has high tracking accuracy and low false alarm rate.	Orientation and length of the Edgelet can't be changed. Not accurate to deal with the harsh environment.	Low resolution cameras provide better timing efficiencies.
Component Based [24]	Detects humans by identifying its body parts like head, legs, and hands separately. Then combine these components using some geographic configuration.	It can be used with cameras having a partial view of the human body.	It is a view-invariant technique. It can deal with noise, occlusion and different lighting conditions.	Detection of each component of a human and checking geometric configuration is a complex and time-consuming process.	Requires higher computational cost than HOG

exist in close proximity. It is capable of handling arbitrary distributions also. Jones and Rehg [77] also used KNN classifier for face detection based human detection.

Adaptive Boosting (AdaBoost) is a machine learning meta algorithm that is suggested by Freund and Schapire [53]. The AdaBoost algorithm can be deployed in aggregation with different machine learning algorithms. It helps to enrich the performance of the classifiers. Huang et al. [73] deployed the AdaBoost learning algorithm to boost the performance of the SVM classifier. Cascade classifier [33, 53, 136] improves the speed of the detection process by uniting the complex classifiers in the multilevel structure, which is a specific instance of ensemble learning. Setjo, Achmad and Faridah [143] used a cascade classifier to identify the existence of humans on the thermal image. Voila and Jones [166] and Aftab Ahmad et al. [2] also deployed cascade classifier for object detection.

Artificial Neural Network (ANN) [18, 20, 89] is an evolutionary approach for information processing encouraged by the biological nervous system. ANN is also be referred to as a connectionist system. It has a remarkable ability to learn patterns from input data and identify the trends based on its learning experience. A neural network has a number of computing elements called neurons. These neurons are connected with each other and organized in layers, which help to convert the input to some output. The internal structure of the neural network can be changed based on the information flowing through the network. This is accomplished by regulating the weight of synapsis or connection. Aibinu, Shafie and Salami [3] have also deployed an artificial neural network for skin detection framework. Zhang et al. [186] used ANN to predict the posture of the human body. Hani K. et al. [5] used a multilayer perceptron (MLP) for skin color classification.

2.4 Deep learning based human detection

With the recent growth of high-performance GPUs and the widespread availability of image data, training for deep learning based models have become more enriched, which introduced several progressive CNN based methods to boost up the object detection performance. The advancement seen for object detection is mainly based on deep neural networks due to its robust feature representation capability. The contemporary deep learning based approaches can detect a single object or multiple objects within an image more accurately than traditional object detection approaches. The human detection framework can accurately detect humans by taking advantage of modern object detection techniques. The recent/modern deep learning based object detectors are briefly described as follows:

2.4.1 CNN

Convolutional neural network (CNN) [1, 10, 13, 92, 95, 115, 152], also known as Covnet, is a class of deep neural networks. The structure of a convnet is similar to the connectivity pattern of neurons in the human brain and was motivated by the visual cortex's organization. It takes an image as input, specifies significance based on learnable weights and bias to several objects in the image, and distinguish one from the other. CNN is generally applied to analyse visual imagery. It usually consists of convolutional layers, pooling layers, fully connected layers and normalization layers. The convolutional layer extracts spatial structure to confer the local representations, while pooling is a shift-invariant operation. AlDahoul et al. [6] developed a convolutional neural network to detect the presence of human being in captured video sequences. They used three diverse deep models named supervised CNN (SCNN), pre-trained CNN and hierarchical extreme learning machine (HELM). HELM [159] is an unsupervised feature learning framework. It delivers more robust features by utilizing the sparse auto-encoders. These deep models are robust against viewpoints, scale, orientations, scale and activities.

The advancement of CNN has led to significant growth in enhancing the performance of different object detectors. The CNN based object detector can be classified into two categories: single-stage and double-stage. Single-stage based object detector consists of one-step framework, which used single or multi-layer sensory neural networks [187] to predict object locations and corresponding class labels directly. In contrast, the double-stage based object detector carries the two-step framework. The first step extracts a set of region proposals using Region Proposal Network [132], or Selective Search [164] or etc. The second step applies CNN to generate the object location and desired class label. The modern object detection techniques are characterized on the basis of the region proposal stage. R-CNN, Fast R-CNN, and Faster R-CNN, Mask R-CNN and Light Head R-CNN follow double stage pipelining, while YOLO v1, v2, v3, and SSD, Tiny SSD, F-YOLO fall under single stage pipelining.

2.4.2 R-CNN

Ross Girshick et al. [60] in 2014 proposed the region based convolution neural network to detect and localize the object within an image correctly. The traditional approaches require a large number of regions for object classification, which spends lots of time. But, R-CNN sidesteps this issue by selecting only 2000 regions from the image. The Fig. 11 shows the overall process involved in R-CNN. RCNN takes a color image as input. It then uses the Selective Search algorithm to extract 2k region proposals from the image. In the next stage, each region proposals are perverted to a square-shaped block and fed into CNN, which generates 4096 feature vectors. Further, these feature vectors are used to build an SVM classifier to categorize the object type included in the corresponding region proposal and bounding box regressor for localization.

The R-CNN requires a large time to train the network and also requires much more disk space to store the feature maps. Hence, it is a time-consuming process.

2.4.3 Fast R-CNN

To deal with several limitations of R-CNN, Ross Girshick et al. [59] in 2015 proposed an improved framework for better object detection and named Fast R-CNN. Figure 12 illustrates the workflow of Fast RCNN. The algorithm of fast R-CNN is quite similar to R-CNN. R-CNN feeds the region proposals to the convolutional neural network. However, Fast R-CNN directly supplies the input image to the Convolutional Neural Network and evaluates essential features. Further, region proposals are identified from these convoluted feature

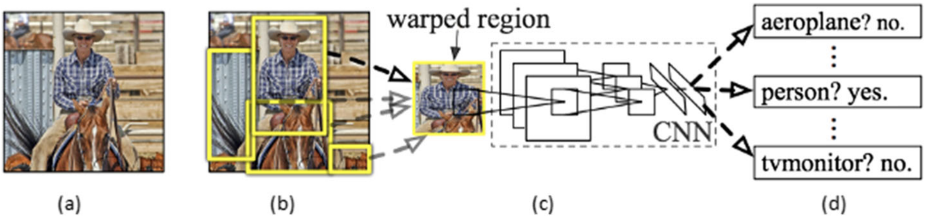


Fig. 11 R-CNN [60] : **a** input image, **b** Extracts region proposals ~(2k), **c** Compute CNN features, **d** Classify regions

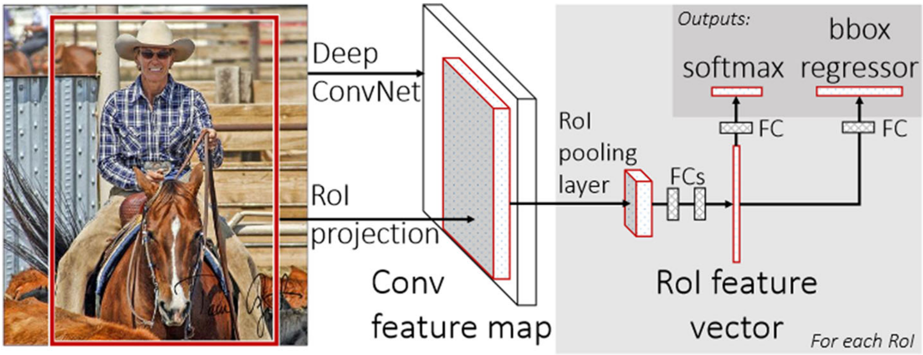


Fig. 12 Fast R-CNN [59]

maps. The ROI pooling layer is used to reshape these proposals into fixed-size blocks so that it can be fed into a fully connected layer. Finally, the softmax and regression layers use ROI feature vectors to predict the class label and offset values for the bounding box.

The Fast R-CNN is quiet faster than R-CNN because the convolution process is performed only once for each image to generate the feature maps. One of the limitations of fast RCNN is that it uses selective search region proposal algorithm to propose regions that slow down performance of the algorithm.

2.4.4 Faster R-CNN

Due to the bottleneck architecture of Fast R-CNN, Shaoqing Ren et al. [132] in 2015 proposed an improved framework for locating objects in real-time environments, named Faster R-CNN. The faster RCNN uses a separate network for region proposals named Region Proposal Network (RPN) algorithm instead of the Selective Search algorithm, which makes it more faster than Fast RCNN. Figure 13 presents the overall workflow of Faster RCNN. Like Fast RCNN, Faster RCNN takes an image as input and passes it to CNN, which produces convoluted feature maps. Further, RPN uses these feature maps to extract the anchor boxes, and these anchor boxes are reshaped into a fixed-sized square using the ROI pooling layer. Finally, these reshaped features are passed to a fully connected layer and then softmax and regression layers are used to predict the class labels and offset the values for bounding box. Faster R-CNN processes an image in 0.2 seconds, and it is 10 × faster than Fast R-CNN and 250 × than R-CNN.

2.4.5 SSD

Liu et al. [96] in 2015 suggested a method to detect objects in real-time within an image by implementing a feed-forward deep neural network, called Single Shot MultiBox Detector (SSD). Faster R-CNN employs an RPN algorithm for generating boundary boxes and use these boxes to classify objects. In contrast, SSD accelerates the detection process by dropping the RPN region proposals algorithm and utilizes little advances, which includes multiscale features and default boxes. This advancement allows SSD to match the Faster R-CNN's accuracy. Figure 14 illustrates the SSD architecture, which consists of three components: Base network, extra feature layers, and prediction layers. The base network is created over the vulnerable VGG-16 design model. Instead of taking VGG fully connected

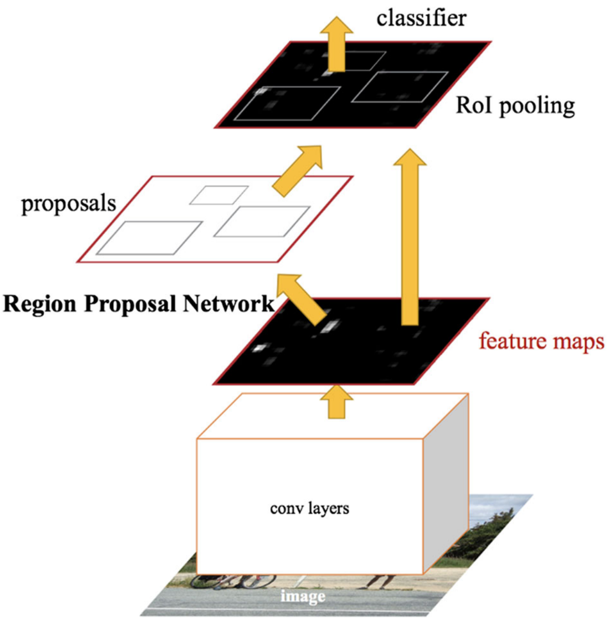


Fig. 13 Faster R-CNN [132]

layers, SSD adds a set of auxiliary convolutional layers or extra features layers. These additional features layers effectively extract meaningful information at multiscale and gradually shrink the input's size to each consequent layer. Prediction layers customize several feature maps for bounding box coordinates to predict classification scores and deal with objects of different sizes. SSD delivers better performance in comparison to Faster RCNN.

2.4.6 YOLO

YOLO (You Only Look Once) is a single-stage detector, developed by Joseph Redmon et al. [131] in 2015. This uses a convolutional network to provide bounding boxes and class probabilities for these boxes. Figure 15 presents the work pipelining of the YOLO detector. This algorithm takes an image as input and split it into the grid of size SxS. The bounding

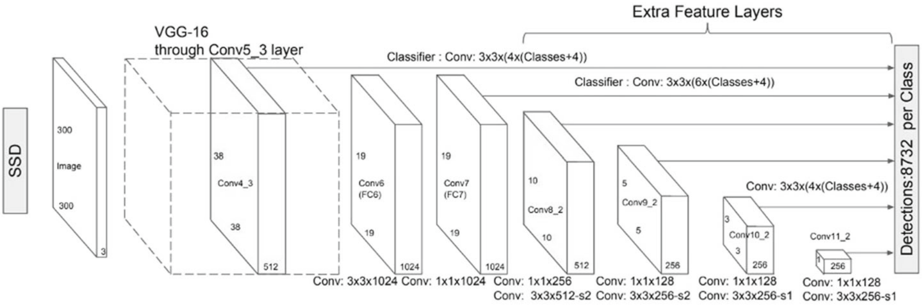


Fig. 14 SSD: Single Shot MultiBox Detector [96]

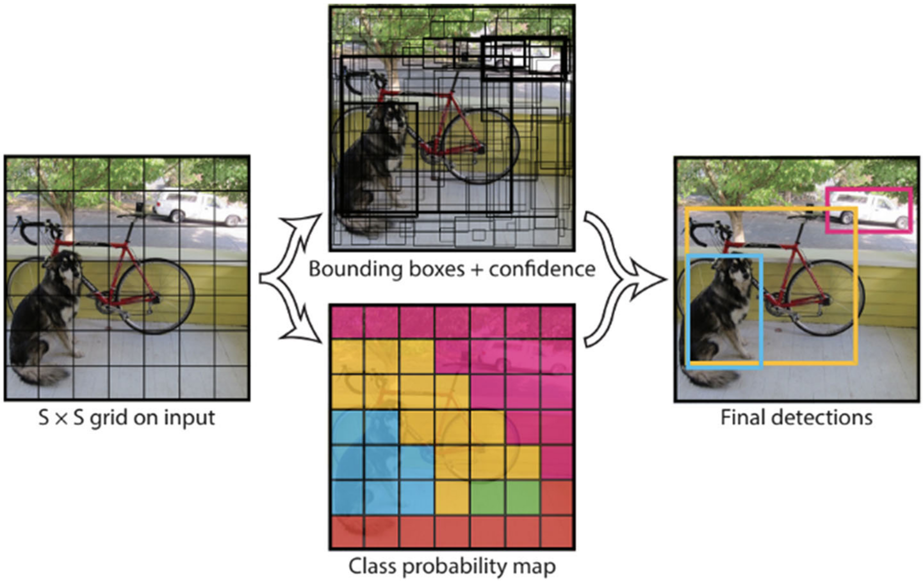


Fig. 15 YOLO: You Only Look Once [131]

boxes of each sized 'm' are taken for each of the grid. The convolutional network yields a class probability and offset values for locating the bounding box as output. The bounding boxes having a higher value of class probability than the predefined threshold value are used to determine the object within an image. It is found that Yolo is faster than R-CNN, Fast R-CNN, and Faster R-CNN due to its faster orders of magnitude (approximate 45 FPS).

In 2017, Joseph and Ali [129] proposed an enhanced framework for detecting 9k object categories, called **YOLOv2** or **YOLO9000**. They developed an advanced methodology to efficiently identify and locate objects with progressive precision and recall values, which makes YOLOv2 faster and more reliable than YOLO v1. YOLOv2 undertakes to utilize the concept of anchor boxes rather than taking k anchor boxes. It searches the most suitable anchor box shapes to locate objects easily. YOLOv2 uses Darknet-19 for feature extraction. It yields 91.2% accuracy and can process the frames with 67 FPS. High-resolution classifier, fine-grained features, and multiscale training make YOLOv2 faster. YOLOv2 learns from the detection-labeled data that concerning the prediction of bounding box coordinate as well as objectness score. From classification-labeled data, YOLOv2 learns new classes of deeper range to extend the number of classes that the model can identify. Finally, it can be said that YOLOv2 is a ground breaking methodology that links the gap between the object classification and detection systems.

In 2018, Joseph and Ali [130] proposed **YOLOv3** to detect objects within the image accurately, which is an incremental version of YOLOv2. It is slightly larger because Darknet-53 was used for feature extraction instead of Darknet-19, but it is more accurate. YOLOv3 uses multiscale prediction to create more vital and finely-grained tuned information from up-sample features and earlier feature maps, respectively. It uses logistic regression for predicting the objectness score of each bounding box and uses binary cross-entropy loss for the class predictions. YOLO v1 and v2 encountered a problem for detecting small objects, but YOLOv3 has performed better to deal with this issue. It is a faster as well as more accurate object detector than others.

2.4.7 Mask R-CNN

Kaiming He [71] proposed a deep neural network based Mask RCNN Network to resolve the instance segmentation problem. Mask R-CNN is a two-stage framework that distinguishes diverse objects within an image. The first stage generates region proposals, and the second stage predicts object classes, bounding boxes, and masks. Mask R-CNN uses Feature Pyramid Networks (FPN) for object detection, which preserves robust semantically features at different scales of resolution. The mask R-CNN can detect objects within an image more accurately than others but is slower.

Apart from these detectors, **Light Head R-CNN**, **Tiny SSD**, and **F-YOLO** are also the best suited real time object detectors. Light Head R-CNN object detector is a two stage approach which was proposed by Zeming Li et al. [93] in 2017. The architectural design of the Light Head methodology creates the network head as lighter as possible. It adopts a small feature map and an economical R-CNN subnet, which is faster and more accurate than the other two stage based approaches. In 2018, Alexander Wong et al. [173] has suggested a new object detector called Tiny SSD, which is a single-shot detection based methodology using CNN for real-time embedded object detection. Tiny SSD used SSD based auxiliary convolutional feature layers to lessen the model size while performing object detection. Its model size is 2.3MB, which is ∼26X lesser than Tiny YOLO and gives ∼4.2% higher mAP than Tiny YOLO. Tiny SSD is well suited for real-time object detection in various embedded scenarios. In 2017, Rakesh and Cemalettin [106] introduced Fast YOLO, which can process hundreds of images in a seconds. The customized network architecture, and distillation loss function make this framework faster. It uses Tiny-Darknet as its baseline method. The customized network of F-YOLO has ten times fewer parameters than a VGG based detector. It can process the test images at the rate of 200 FPS.

The comparative analysis of deep learning based techniques is described in Table 4. This table compares different techniques based on various characteristics such as the key method used, the concept behind the object detector, its pros, and cons. Whereas Table 5 describes the real-time performance of these detectors based on various attributes such as input image size, employed backbone architecture, mAP (mean average precision) values obtained for VOC and COCO dataset, processing speed in FPS, and remarks. The detectors in the R-CNN family follows two-stage pipelining, while Yolo and SSD family follows one stage pipelining. R-CNN is a popular and efficient object detector but cannot be used in real-time scenarios because it requires huge computational time to a process test image. Light Head R-CNN is the fastest algorithm in the R-CNN family and can process up to 102 frames in a second. However, Fast-YOLO has achieved a new milestone among different object detectors by processing up to 200 frames in a second.

3 Human detection dataset and performance measure

3.1 Human detection datasets

In the most recent decades, numerous datasets have been made freely accessible for assessment of human detection frameworks. These datasets (illustrated in Table 6) can be utilized as a benchmark for different scope of applications like video surveillance, CBIR, service

Table 4 Comparative analysis of deep learning based techniques

Object Detector	Year	Region-proposal/Key method	Concept	Pros	Cons.
CNN [142]	2014	Sliding Window	First, the image is partitioned into different regions, and then each region is classified into separate classes.	It can detect as well as locate single or multiple objects within an image proficiently.	Requires high computational cost
R-CNN [60]	2014	Selective Search	It uses Selective Search algorithm for region proposals. CNN and SVM are used for feature extraction and object classification purposes, respectively.	Instead of working with the massive number of regions, it extracts ~2k regions and checks the existence of the object in these regions..	It requires large computation time and cannot be implemented in real-time because it takes ~47 seconds to process the test image.
Fast R-CNN [59]	2015	Selective Search	Feature maps are extracted first using CNN, and then Selective Search is used to generate region proposals. Softmax and linear regression are used in the top of the fully connected layer to predict object class and bounding box coordinates.	Instead of feeding 2k region proposals to CNN, it feeds the input image directly to CNN and generates feature maps, which makes this algorithm faster than R-CNN.	Selective search is slow and time-consuming process. It takes 2 seconds to process the test image.
Faster R-CNN [132]	2015	Region Proposal Network	Faster RCNN is similar to fast-RNN, but it used Region Proposal Network in place of the Selective Search algorithm.	It used Region Proposal Network for extracting regions, which accelerates the performance of faster RCNN. It takes 0.2 seconds to process the test image.	Since different systems are tuned to single pipelining, the performance of the systems depends on the performance of the previous systems, so it takes time for object proposal.
YOLO v1 [131]	2015	Unified Model* (consists of 24 convolutional layers followed by 2 fully connected layers)	YOLO uses single convolutional network to evaluate multiple bounding boxes and corresponding class probabilities. Further, it thresholds the resulting detections by the model's confidence for accurate detection.	Having a higher order of magnitude than R-CNN families; less likely to predict false positives on the background.	Not much accurate than many detectors; unable to deal with small objects.
SSD [96]	2016	Multiscale features and default boxes	SSD uses vulnerable VGG16 design as the base model. Multi-scale feature maps and default bounding boxes are used to detect the objects independently.	SSD is faster than RCNN families because it takes one single shot to identify multiple objects within the image.	It is also struggling to detect small objects.

Table 4 (continued)

Object Detector	Year	Region-proposal/Key method	Concept	Pros	Cons.
YOLO v2 [129]	2016	Anchor Box	YOLOv2 uses Darknet-53 for feature extraction. The fined grained features and Multi-scale training are used for object detection.	Achieved high accuracy and speed; It can detect up to ~9k objects. It takes 0.14 seconds to process a test image.	Not suitable for identifying tiny objects especially when they are in group.
Mask R-CNN [71]	2017	Feature Pyramid Network	It extends the Faster R-CNN by uniting a branch to predict the mask of an object parallel with the actual branch for bounding box detection.	It solved the instance segmentation problem. It can accurately detect various objects in an image.	Not as fast as Faster-RCNN
Light Head R-CNN [93]	2017	Region Proposal Network	Light Head R-CNN uses a thin feature map and cheap R-CNN subnet for making the head of network as lighter as possible.	The lighter head of the network, united with a tiny Xception-like base model, makes it faster and takes ∼ 0.008 seconds to process the test image.	It is two stage detector, in which one stage is totally depends on another stage.
YOLO v3 [130]	2018	Anchor Box	YOLOv3 uses Darknet-53 for feature extraction. Multiscale prediction is used to create more vital and finely-grained tuned information. Logistic regression and binary cross-entropy loss are used to predict objectness score and class prediction, respectively.	Provides accurate bounding box and class prediction; achieved high accuracy over YOLOv2; It takes 0.022 seconds to process a test image.	Not as fast as YOLOv2 and still facing issue when identifying small objects in groups.
Tiny SSD [173]	2018	Optimized Fire Sub-Network Stack	Tiny-SSD uses optimized SqeezeNet and optimized sub-network stack of SSD based convolutional feature layers for faster object detection.	Its model size is 2.3MB, which is ~26X lesser than Tiny YOLO and gives ~4.2% higher mAP than Tiny YOLO.	-
F-YOLO [106]	2018	Compact network architecture based on Tiny-YOLO and distillation loss function	It uses Tiny-Darknet as its baseline method. F-YOLO applies distillation loss function on a single pass detector to improve the performance; Passes the knowledge from convoluted feature maps of the teacher network to the proposed student network.	It is light weight, single stage, much faster object detector. It has fewer parameters than VGG based detectors. It takes ~0.005 seconds to process the test image.	Faster but not as accurate as others.

Table 5 Performance analysis of CNN based object detectors

Object Detector	Input Image Size (w, h)	Backbone Architecture	mAP(%)			FPS	Remarks
			VOC 2007	VOC 2012	COCO		
R-CNN	(227, 227)	AlexNet	58.5	53.3	–	<0.1	Reduced region proposals size to ∼2k.
Fast RCNN	∼(1000, 600)	VGG-16	70	68.4	19.7	0.5	It is 9 × faster than R-CNN at train-time and 213 × faster at test-time.
Faster RCNN	∼(1000, 600)	Resnet-101	78.8	70.4	27.2	7	Faster and accurate than RCNN, Fast RCNN
YOLO v1	(448, 448)	GoogleNet*	63.4	57.9	–	45	Faster and having unified architecture
SSD	(300, 300)	VGG16	77.2	75.8	23.2	46	Can distinguish more than just 200 object categories
YOLO v2	(544, 544)	Darknet-19	76.8	73.4	21.6	67	Can distinguish more than 9k object categories.
Mask RCNN	∼(1000, 600)	FPN	–	–	35.7	5	Accurate but less faster than Faster R-CNN
YOLO v3	(320, 320)	Darknet-53	–	–	28.2	45	Having better detection ability at different scales
Light Head	(1100, 700)	Xception*	–	–	30.7	102	Fastest object detector in R-CNN family
Tiny SSD	(300, 300)	SqueezeNet	61.3	–	–	–	Process 2.3MB model size
F-YOLO	(416, 416)	Tiny Darknet	59.4	–	–	200	Extreme faster; Having 10 times fewer parameters than a VGG based object detector

Table 6 Human detection datasets

Dataset Name	Training Records/ Testing Records	Video/ Image	Color/ Grey	Occlusion	Scenes	Comment
MIT [31, 166]	Total number of images are 925.	Image	Color	–	Outdoor	Created in 2000. The compressed size of the dataset is 10 MB. Resolution=64×128 PPM format images
CAVIAR [31, 40, 65]	People walking alone class contains video sequences with size mostly between 6 MB to 12 MB at 25 frames/sec.	Video	Color	Inter object	Outdoor	Created in 2004. Videos are captured at different viewpoints and illuminations with 384×288 resolution.
INRIA [31, 54, 68, 73]	Consist of 2416 positive and 1218 negative images for training. Consist of 1127 positive and 453 negative images for testing.	Image	Color	–	Outdoor	Created in 2005. This dataset contains static images and incorporated variations in cropped humans with 64×128 resolution.
PASCAL VOC [49]	Consist of 20 classes in which 11530 images are from train/validation having 27450 ROI objects annotation and 6929 segmentations.	Image	Color	Inter object & self	Outdoor	Created eight variants from 2005 to 2012.
DC [111]	Include 24k images of humans and 39k images of non-humans	Image	Grey	–	Outdoor	Created in 2006
CVC [31, 63, 168]	Six variants of CVC (CVC 01 to CVC 06) from 2007 to 2013 are created with a diverse environment. Images are captured through Visible and FIR cameras.	Video	Color	Inter object	Outdoor	This dataset includes annotated pedestrian and also incorporated partial occlusion.
Penn-Fudan [169]	Contains 345 labelled pedestrian from 170 images.	Image	Color	–	Outdoor	Created in 2007. The height of labelled pedestrian fall into 180×380 pixels.
TUD Pedestrians [7]	Consist of 250 images with 250 fully visible persons.	Image	Color	Inter Object	Outdoor	Created in 2008. Images are having variations in articulation and different clothing.
NICTA [105]	Consist of 187000 positive and 5200 negative images for training. Consist of 6900 positive and 50000 negative images for testing.	Video	Color	–	Outdoor	Created in 2008. Some challenging conditions are included for validation purpose.

Table 6 (continued)

Dataset Name	Training Records/ Testing Records	Video/ Image	Color/ Grey	Occlusion	Scenes	Comment
TUD-Brussels [72]	Consists of 1092 positive and 192 negative records for training. Consist of 508 testing records.	Video	Color	Inter Object Occlusion	Outdoor	Created in 2009. It has been created under the urban environment.
PETS 2009 [51, 114, 168]	Consist of one training set for people tracking with different viewpoints and different video size	Video	Color	–	Outdoor	Created in 2009. This dataset contains inter-object occlusion. The resolution of videos is 768×576.
CALTECH [114, 168]	Consist of 6 training sets with each having 6-13 one minute long video file. Consist of 5 testing sets.	Video	Color	Inter object	Outdoor	Created in 2009. Total size ≈ 11GB (compressed). Annotation file for the entire dataset is available. Comprehensive recognition outcomes are also incorporated.
CHUK [116, 168]	Consists of 10 clips and incorporates 1063 images with occluded pedestrians from different datasets.	Video	Color	Inter object	Indoor	Created in 2012. Ground truths are labelled manually for all pedestrians.
HDA Person [112]	Total 75207 frames with involving more than 80 persons.	Video	Color	Inter object	Indoor	Created in 2013. This dataset involves non-occluded and occlusion person.
Daimler Pedestrian Path Prediction Benchmark [137]	It contains sequences having 19612 stereo image pairs in which 12485 images are manually labelled pedestrian bounding boxes and 9366 images comprising pedestrian detector quantities.	Video	Grey	–	Outdoor	Created in 2013. Sequences are made with one pedestrian and there is no occlusion.
PETA [40]	Total images are 19000 in which 8705 are persons.	Image	Color	Inter object and self	Varying	Created in 2014. Incorporated variations in Scenes. Resolution of images is ranging from 17-by-39 to 169-by-365 pixels.

assistance, etc. and these are gathered from a wide range of scenarios. Some of the general-purpose datasets for human detection are USC-A, Human Eva, CMU, USC-C, INRIA, MIT, PASCAL-VOC, PENN-FUDAN, and H3D datasets [24, 49, 72, 73, 113, 168]. The CAVIAR and USC-B datasets are used in surveillance applications. The Miao, Daimler-Chrysler, CVC, Caltech and TUD datasets [7, 31, 63, 72, 137, 168] can suitably be used in the application of pedestrian detection. These datasets consist of several complexities with the form of human-like occlusion, appearance, viewpoint, pose etc. For example, the frontal view of humans has demonstrated in USC-A and MIT datasets. The several poses and viewpoints are illustrated in H3D, PASCAL, USC-C, INRIA, VOC and Penn-Fudan datasets. The occlusion for humans might be possible in H3D, PASCAL, USC-B, VOC, Caltech and CAVIAR datasets [54, 114, 161, 168]. In the CAVIAR and USC-B datasets, there are some footage videos in which occlusion between the humans are present. The different cases of partially occluded pedestrians are present in Caltech datasets. Sometimes, unappropriated image borders, camera viewpoints, human interaction and non-human objects may also have introduced the occlusions. H3D, VOC and PASCAL datasets are having this kind of occlusion.

3.2 Performance measures

Assessment of the performance of human detection framework is a foremost task. Its purpose is to homologate the robustness and correctness of the system. The assessment of measures for the human detection system can be possible in a qualitative and quantitative manner. The evaluation of qualitative approach is based on visual analysis, which comprises various issues. On the other side, the quantitative assessment approach needs ground truth data to compute the outcome using numeric assessment, which is a highly challenging task. The number of measures [8, 21, 32, 39–42, 63, 66, 88, 146, 148, 161, 162, 170, 177] like precision, recall, accuracy, etc. is described in this section to evaluate the performance of the human detection system in a quantitative manner.

Contingency table It is also known as Error matrix or Confusion matrix [7, 55, 57, 79, 103, 109, 113, 117, 121, 137, 140, 160, 163, 168, 176, 183]. It is a table that is used to measure the performance of supervised learning algorithm. It is a summary of predicted outcomes. This is calculated over the test dataset in which ground truth values are known. In the confusion matrix, columns denote the predicted class and rows denote the actual class. It is used to measure the different quantitative parameters like precision, recall, accuracy, f-measure, etc.

Accuracy Accuracy [11, 81, 150] evaluates the correctness and wellness of the classification model. It specifies how close an outcome approaches the actual value. Accuracy is a ratio of the number of accurate predictions to the total number of samples.

$$Accuracy(\%) = \frac{(T_p + T_n)}{(T_p + T_n + F_p + F_n)} \tag{36}$$

Where T_p, T_n, F_p, and F_n are true positive, true negative, false positive and false negative respectively. These are described as follows in a general way:

True Positive: The model predicted Yes and actually, it was Yes.

True Negative: The model predicted No and actually, it was No.

False Positive: The model predicted Yes and actually, it was No.

False Negative: The model predicted No and actually, it was Yes.

Precision [P] Precision [55, 113, 121, 146, 148, 149] represents the relevant samples only. It is a ratio of correctly classified relevant samples to the total number of predicted relevant samples. The higher value of precision signifies that there is less number of false positive and samples are correctly classified.

$$Precision(\%) = \frac{T_p}{(T_p + F_p)} \tag{37}$$

Recall [R] Recall or Sensitivity or True Positive Rate [55, 113, 121, 146, 148, 149] represents all the relevant samples. It is the ratio of correctly classified relevant samples to the total number of all relevant samples. The higher value of recall signifies that there is less number of false negative and class is correctly recognized.

$$Recall(\%) = \frac{T_p}{(T_p + F_n)} \tag{38}$$

F-Measure [F] F-Measure [55, 62, 113, 148, 184] evaluates the test accuracy by Harmonic mean of precision and recall. It shows the correctness (How many samples are classified correctly) and robustness (It doesn't miss the significance amount of samples) of the classifier.

$$F(\%) = (\frac{2 \times P \times R}{(P + R)}) \times 100 \tag{39}$$

Specificity Specificity or True Negative Rate [55, 74, 113, 186] calculates the proportion of actual irrelevant samples that are appropriately classified.

$$Specificity(\%) = \frac{T_n}{(T_n + F_p)} \tag{40}$$

PED and PAT scores PED score [62, 113] refers to the percent event detected score, which denotes the ratio of the number of real alarms detected in the ground truth to the total number of alarms in ground truths. PAT score [62, 113] refers to the percent alarms true score, which denotes the ratio of the number of real alarms detected in the ground truth to the total number of alarms detected by modules. The higher value of PAT score specifies that the false alarm is rarely triggered by the module. While the higher value of PED score signifies that most of the objects are detected that should start an alarm.

ROC Curve ROC Curve [62] refers to the receiver operating curve, which is a graphical plot that illustrates the performance of the classification model. It is constructed over true positive rate against false positive rate at diverse threshold values.

4 Comparison of different human detection techniques

So far, several techniques of human detection have been proposed by various researchers. These human detection techniques can be purely image processing based or feature learning or deep learning paradigm based. The purely image processing based human detection frameworks are straightforward and faster because they deal directly with pixels to detect humans' presence. Skin color modeling, head curve geometry, frame differencing, and optical flow algorithm comes under this category. These techniques may be affected by various environmental factors such as inadequate lighting conditions, reflection, cluttered background, moving objects in the background, etc., which lead to incorrect results. In the next type, feature learning based human detection frameworks become more popular because they can detect humans more accurately than purely image processing based frameworks. Feature learning based frameworks follow two-stage pipelining, in which the first stage extracts the features, and the next stage uses these features to create classifiers. Viola-Jones algorithm is a popular face feature based human detector that is 15 times faster than skin color modeling based techniques. Deep learning-based human detectors herald a new revolution in the field of object detection. There is no requirement of any additional feature extraction techniques like feature learning based framework. These detectors extract features implicitly with the help of deep convolutional neural network. The deep learning based detectors require massive computation in training time. The detectors can locate single or multiple humans in an image more accurately than others.

This section briefly summarizes the comparative analysis of various researches on human detection. Table 7 outlines the different human detection strategies based on different parameters. These parameters include the research reference with proposed year, purpose of the research, methods used, classification model used, databased used, training & testing module information, pros, and results. Based on the cited papers, this table contains the results of detection in terms of performance measures such as accuracy, F-measure, TPR, FPR, FPS, etc.

On analysis of Table 7, it has been observed that many techniques or modalities require a classifier to detect the human. Skin color modeling based techniques are fast and computationally efficient, but their performance is dependent on color space. Voila jones [166] has proved to itself the most effective framework for face detection because its computational cost is fifteen times faster in comparison to other traditional approaches. Fu-Chun Hsu et al. [75] have experimentally proved that $HOOF + SVMclassifier$ performed better than $HOG + SVMclassifier$ and $HOG + HOOF + SVMclassifier$. The combination of $HOOF + SVMclassifier$ is capable of dealing with low-resolution videos and occlusion. The geometry based method, proposed by Wong and Hui [175], can detect only a single human at a time.

Han and Tong [68] have used geometric flow based bandlet transform for real-time human detection. They found the accuracy of 97% at 0.2 FPR. The methodology, proposed by Bhaskar Chkraborty et al. [24], can detect humans in the cluttered environment and can also deal with crowded scenes. Seemanthini and Manjunath [140] have proved that their proposed method can deal with different objects of diverse sizes and shapes. The framework proposed by Vijay and Shashikant [54] has used edgelet features for detecting the human with accuracy 95% at 1.1 false alarm rate. D.-S. Huang et al. [73] suggested a good methodology for human detection. This algorithm is capable of detecting the human in occlusion and non-occlusion environment efficiently with high detection rate at low false positive rate. In addition to feature based approaches, intensive learning based techniques for human detections have also been briefly summarized, indicating an intuitive improvement in the field of human detection. Chahyati et al. [22] used CNN facilities to track multiple individuals for surveillance purposes. This proposed framework adopted Faster R-CNN for detection purposes and employed two different methods (such as Euclidean distance and Simense neural network) for object association. In experimentations, it is found that Euclidean distance provides encouraging outcomes as an object association method, but it gradually depends on the individual frames' detection process. Agus, Rohmat, and Purwono et al. [110] suggested an advanced framework for the driver assistance system using low constraint devices

Table 7 Comparative analysis of different human detection strategies

Author & Year	Purpose	Technique Used	Classification Model	Database Used	Training and Testing	Pros.	Result
Maheswari and Reeba [102] in 2017	Skin Tone Detection	Color Space: YCbCr model, color Thresholding	No	NO	No	High Accuracy. This model can also be deal with biomedical image processing.	Average Detection Rate=93%
Hani K. et al. [5] in 2015	Hybrid Human Skin Detection	YIQ color space, Skin Color and Texture descriptor	MLP-ANN + Kmeans Clustering	ECU Database	1,173,435 skin pixels out of 2,400,000 total pixels are used to build the classifier.	Capable to generalize any kind of specified data with the help of neurons.	F1 measures = 87.82%
Viola and Jones [166] in 2001	Human Face detection	Haar Feature Selection, Integral Image, Adaboost Learning Algorithm	Cascade Classifier	MIT + CMU Dataset	4916 face image and 9544 non-face image are used to build the classifier at base resolution 24x24.	15 times faster than previous approaches.	Accuracy up to 93.9%
Aftab Ahmed et al. [2] in 2018	Face Recognition at low resolution	LBPH	Cascade Classifier	LR500	Kept 500 images for each person, which helps to build classifier.	Detects faces in various angles, side poses and also track the human face while human in motion.	Accuracy up to 94%
Rusia, Singh and. Ansari [134] in 2019	Human Face detection and recognition	LBPH	Cascade Classifier	Self-Created	Consist 5 classes, total 750 images for training and 250 images for testing	Best suitable for grey scale image	Accuracy up to 81.6%
Wong and Hui [175] in 2012	Face detection by using thermal image	Pre-processing, point searching and curve drawing	No	No Dataset used	No training Module	It can deal with thermal image and can be used for surveillance in dark.	Overall Accuracy= 91.4%

Table 7 (continued)

Author & Year	Purpose	Technique Used	Classification Model	Database Used	Training and Testing	Pros.	Result
Jiajia Guo et al. [65] in 2017	Moving Object Detection	Combination of frame differencing and background subtraction	No	CAVIAR Dataset	No Training Module.	It also reduces the noise and fills the cavity efficiently.	This proposed method can detect moving object efficiently in comparison to other traditional techniques.
Deepjoy and Sarat [38] in 2014	Human Detection	Background Subtraction	No	No	No Training. Algorithm tested for human walk at slow, medium and fast walk in front of digital camera in outdoor and indoor settings.	It is simple and fast algorithm.	Real time robust background subtraction and shadow detection method performs better than other two.
Kiran Kale et al. [80] in 2015	Moving Object Tracking	Median filter, optical flow, object segmentation, motion vector estimation	No	Traffic Closed Circuit TV and real time videos	No Training Module. Tested this proposed method for multiple event such single human, speed variation, stationary object and multiple object, etc.	Multiple objects in videos can be tracked.	Average accuracy= 91.42
Suman Kumar Choudhury et al. [31] in 2018	Pedestrian Detection	Silhouette orientation histogram, Golden Ratio Based Partition	Linear SVM, HIKSVM	INRIA, CVC-01, CAVIAR, MIT-Traffic datasets	Not Mentioned	Vast analysis has been done. Proposed methodology takes 13ms to process per frame. Time can be reduced by optimization or parallelization.	With varying parameter C for INRIA dataset, Change in accuracy = up to 92.22% (test accuracy) for SVM and up to 92.55 (Test accuracy) for HIKSVM.
Fu-Chun Hsu et al. [75] in 2013	Human Head Detection for crowd monitoring system	Integral HOOF, HOG	SVM	Not Specified	Collected 571 positive and 429 negative samples from 16 video recording	HOOF produced good result than HOG. It can deal to low qualities videos with occlusion.	Overall Accuracies are 65.52%, 88.48% and 86.26% for HOG+SVM, HOOF+SVM, and HOG+HOOF +SVM respectively

Table 7 (continued)

Author & Year	Purpose	Technique Used	Classification Model	Database Used	Training and Testing	Pros.	Result
Han and Tong [68] in 2013	Real time human motion detection	Region segmentation based on optical flow, Bandlet Transform based on geometric flow	SVM	INRIA Human Dataset	From a selection of 6246 images, 2416 positive samples and 1377 negative samples are used to build classifier. 1132 positive samples and 821 negative samples image are used to form test data set.	The method can mark human motion efficiently while resisting the background movement with low computational cost.	Accuracy=97% at false positive rate = 0.2
Greche and Najia [64] in 2016	Facial Expression Recognition	HOG	Normalised Cross Correlation	Cohn Kanade Dataset, CFD Dataset	Not Given, This model classified in to 5 classes: anger, surprise, joy, sadness, neutral. Each image having 64x64 resolution.	Low computation cost with low resolution image. This approach extracts the local feature efficiently.	Inward Comparison Average accuracy 83.09 and 88.11% for CK dataset and CFD dataset respectively.
Seemanthini and Manjunath [140] in 2018	Human detection and Tracking	Cluster Segmentation, Temporal tracking, HOG for feature extraction	SVM	Not Mentioned	Not Mentioned	This approach can deal with different objects of diverse shape and size.	Accuracy 89.59%
Felip and Helio[39] in 2017	Violent Events in Video Sequences	CENTRIST for feature extraction and PCA for dimensionality reduction	SVM	Violence Flows Dataset, Hockey Fight Dataset	For violent dataset, 256 crowded scenes in which 123 for violence scene and 123 for non-violence. For hockey fight, 1000 clips in which 500 for fight scenes and 500 for non-fight scene.	Can be used in thermal / RGB images; Less time complex.	Achieved 90.69% accuracy for hockey fight dataset and 86.16% accuracy for violent flow dataset

Table 7 (continued)

Author & Year	Purpose	Technique Used	Classification Model	Database Used	Training and Testing	Pros.	Result
Vijay and Shashikant [54] in 2015	Pedestrian Detection	Edgelet Features	Cascade structure of K-Means clustering for Classification	INRIA Dataset	Division of training dataset is in to three groups according to human distance from camera : near, medium and far.	Suitable for real time implementation due to its minimum computational complexity.	Detection Accuracy is 95% at false alarm rate of 1.1
Bhaskar Chakraborty et al. [24] in 2007	Human Detection	HOG	SVM	Human Eva Dataset	10000 positive and negative example for head classifier, 104500 positive and 20000 negative example for legs. The number of example is same for arm classifier as legs classifier.	It can handle noise and various lighting conditions. It can detect the human in the cluttered scene.	Not Specified
D.-S. Huang et al. [73] in 2014	Human Detection	HOG + Para-llogram based Haar like features (PHF)	AdaBoost + SVM	INRIA Dataset	Training set comprise of 2500 human and 5000 non-human. Test data comprises 4500 positive and 15000 negative samples	Detection can handle crowded scenes and different illumination conditions.	For occlusion, detection rate is 0.96 at FPR=0.32. For non-occlusion, Detection Rate is 0.99, at FPR=0.9.
Chahyati et al. [22] in 2017	Tracking pedesterians for surveillance	Faster R-CNN, Siamese neural network, Euclidean System	–	Self created dataset	Fourteen scenes are taken from the same area at different times contain 4-8 trajectories from 21 – 89 frames.	Euclidean distance provides encouraging outcomes as an object association method, but it gradually depends on the individual frames' detection process.	Overall accuray= 75%
AlDahoul, Nouar, et al. [6] in 2018	Human Detection	Optical flow	SCNN, CNN, HELM	UCF-ARG aerial dataset	160 videos for training and 80 videos for testing.	HELM doesn't entail iterative fine tuning of weights.	HELM produced 95.9% accuracy.

Table 7 (continued)

Author & Year	Purpose	Technique Used	Classification Model	Database Used	Training and Testing	Pros.	Result
Agus, Rohmat, and Purwono et al. [110] in 2019	Advance Driver Assistance System	SSD	–	VOC 2007	Not mentioned	It tracks humans (pedestrians, cyclists, and riders) in real-time using a low-constrained device such as Raspberry Pi.	Takes 0.8 FPS with 77.6% processor consumption and 70.3% memory when detecting humans.
Kim, Bubryur and Yuvaraj et al. [86] in 2020	Pedesterian detection for smart building surveillance	Optimized VGG-16	–	INRIA	Consisting of 3578 images of pedestrians and 3239 images of non-pedestrians	Fast and accurate	Accuray= 98.5%
Aichun, Tian and Qiao [190] in 2019	Multiple human upper bodies detection	Candidate-Region Convolutional Neural Network(CR-CNN)	–	TV Human Interaction dataset	TVHI dataset contains 300 video clips compiled from 23 different TV shows and annotated with the bounding boxes of human upper bodies.	Multiple convolutional features are included in CR-CNN to produce local and relevant information from the image, showing the effectiveness of the proposed method.	Achieved accuracy up to 86% for TVHI dataset
Mateus, David and Miraldo et al. [104] in 2019	Pedestrian Detection for Human-Aware Navigation	Aggregate Channel Features (ACF) detector	deep Convolutional Neural Network(CNN)	INRIA	Comprises 1832 training images, from which 1218 are negative images, and 614 are positive images; Comprises 288 test images	The proposed framework for robotic navigation applications is quite robust and fast and can process up to 10 frames per second.	Avg. no. det. ACF + CNN = 2.36 with Frame rate=7.85 FPS (for Baseline); avg. no. det. ACF + CNN = 1.81 with Frame rate = 10.41 FPS (for Thresold)

such as Raspberry Pi. This framework used a multiscale object detection model, named Single Shot Multibox Detector (SSD) for human detection. Their experiments showed that the low constraint device takes 0.8 FPS with 77.6% processor consumption and 70% memory when detecting humans. Kim, Bubryur and Yuvaraj et al. [86] used an optimized VGG-16 network to detect pedestrians for smart building surveillance. This method used the INRIA dataset for training purposes and achieved up to 98.5% detection accuracy. Mateus, David, and Miraldo et al. proposed a pedestrian detection framework for human-aware navigation using deep neural networks. They used the Aggregate Channel Features (ACF) detector with deep Convolutional Neural Network(CNN) to develop a fast and efficient object detector. The proposed framework (incorporating both the pedestrian detection and the human-aware constraints) is quite robust and can process up to 10 frames per second. Finally, it is concluded that deep learning based object detectors are more accurate and relatively faster in detecting humans in real-time surveillance than others.

5 Conclusion

For the past few decades, finding the objects in the video sequences has been a consequential and influential research topic in computer vision and video processing. It is really a difficult task to process the low-resolution digital images of any computer vision application. In this paper, we have discussed various human detection techniques in detail. This paper has classified human detection techniques into four categories. These categories are face feature based, motion feature based, shape feature based, and deep learning based human detection. For each technique, their positive aspects & shortcoming have been discussed on the basis of nature of the algorithm. Apart from these, their concept, applicability and real-time performance are also presented in their respective sections. In addition, a comprehensive survey of various state of the art techniques is also presented here. The performance of these algorithms can be evaluated from several parameters like precision, recall, true positive rate, false positive rate, accuracy, f-measures, etc. These measurement parameters evaluate the detection rate and the error rate of the technique. Research is still being done to propose better algorithms. The techniques that are discussed in this paper are good to work in day vision. So, there is a need to develop an efficient algorithm that can deal with the night vision too. The human detection approaches can also be improved more as there is a need to build efficient descriptors that can fulfil the purpose of detection and deal with various occlusion conditions. We expect that this survey article on the development of human detection can help researchers to conduct research related to this field.

References

1. Abughalieh K, Alawneh S (2020) Pedestrian orientation estimation using cnn and depth camera. Tech. rep., SAE Technical Paper
2. Ahmed A, Guo J, Ali F, Deeba F, Ahmed A (2018) Lbph based improved face recognition at low resolution. In: 2018 international conference on artificial intelligence and big data (ICAIBD), IEEE, pp 144–147
3. Aibinu AM, Shafie AA, Salami MJE (2012) Performance analysis of ann based ycbcr skin detection algorithm. Procedia Eng 41:1183–1189
4. Al-Mohair HK, Saleh J, Saundi S (2013) Impact of color space on human skin color detection using an intelligent system. In: 1st WSEAS international conference on image processing and pattern recognition (IPPR'13), vol 2

5. Al-Mohair HK, Saleh JM, Suandi SA (2015) Hybrid human skin detection using neural network and k-means clustering technique. Appl Soft Comput 33:337–347
6. AlDahoul N, Sabri M, Qalid A, Mansoor AM (2018) Real-time human detection for aerial captured video sequences via deep models. Comput Intell Neurosci p 2018
7. Andriluka M, Roth S, Schiele B (2008) People-tracking-by-detection and people-detection-by-tracking. In: 2008 IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8
8. Ansari MA, Dixit M (2017a) An image retrieval framework: a review. Int J Adv Res Comput Sci 8(5)
9. Ansari MA, Dixit M (2017b) A refined approach of image retrieval using rbf-svm classifier. Int J Signal Process Image Process Pattern Recognit 10(9):43–56
10. Ansari MA, Singh DK (2018) Review of deep learning techniques for object detection and classification. In: International conference on communication, networks and computing. Springer, New York, pp 422–431
11. Ansari MA, Kurchaniya D, Dixit M (2018) A comprehensive analysis of image edge detection techniques. Int J Multimed Ubiquitous Eng 12:1–12
12. Astawa INGA, Putra KG, Sudarma M, Hartati RS (2017) The impact of color space and intensity normalization to face detection performance. Telkomnika 15(4):1894–1899
13. Bajaj K, Singh DK, Ansari MA (2020) Autoencoders based deep learner for image denoising. Procedia Comput Sci 171:1535–1541
14. Bhoyar K, Kakde O (2010) Skin color detection model using neural networks and its performance evaluation. In: Journal of computer science. Citeseer, New Jersey
15. Bhuvaneswari K, Rauf HA (2009) Edgelet based human detection and tracking by combined segmentation and soft decision. In: 2009 international conference on control, automation, communication and Energy Conservation, IEEE, pp 1–6
16. Bianconi F, Bello R, Fernández A, González E (2015) On comparing colour spaces from a performance perspective: application to automated classification of polished natural stones. In: International conference on image analysis and processing. Springer, New York, pp 71–78
17. Brand J, Mason JS (2000) A comparative assessment of three approaches to pixel-level human skin-detection. In: Proceedings 15th international conference on pattern recognition. ICPR-2000, IEEE, vol 1, pp 1056–1059
18. Brown DA, Craw I, Lewthwaite J (2001) A som based approach to skin detection with application in real time systems. In: BMVC, vol 1. Citeseer, New Jersey, pp 491–500
19. Burger W, Burge MJ (2016) Colorimetric color spaces. In: Digital image processing. Springer, New York, pp 341–365
20. Bush IJ, Abiyev R, Ma'aitah MKS, Altıparmak H (2018) Integrated artificial intelligence algorithm for skin detection. In: ITM web of conferences, EDP sciences, vol 16, p 02004
21. Cai J, Goshtasby A (1999) Detecting human faces in color images. Image Vis Comput 18(1):63–75
22. Chahyati D, Fanany MI, Arymurthy AM (2017) Tracking people by detection using cnn features. Procedia Comput Sci 124:167–172
23. Chai D, Ngan KN (1999) Face segmentation using skin-color map in videophone applications. IEEE Trans Circ Syst Video Technol 9(4):551–564
24. Chakraborty B, Rius I, Pedersoli M, Mozerov M, Gonzàlez J (2007) Component-based human detection
25. Chandrappa D, Ravishankar M, RameshBabu D (2011) Face detection in color images using skin color model algorithm based on skin color information. In: 2011 3rd international conference on electronics computer technology, IEEE, vol 1, pp 254–258
26. Cheddad A, Condell J, Curran K, Mc Kevitt P (2009a) A new colour space for skin tone detection. In: 2009 16th IEEE international conference on image processing (ICIP), IEEE, pp 497–500
27. Cheddad A, Condell J, Curran K, Mc Kevitt P (2009b) A skin tone detection algorithm for an adaptive approach to steganography. Signal Process 89(12):2465–2478
28. Chen N, Chen WN, Zhang J (2015) Fast detection of human using differential evolution. Signal Process 110:155–163
29. Chen Y, Tian Y, He M (2020) Monocular human pose estimation: a survey of deep learning-based methods. Comput Vision Image Understand 192:102–897
30. Chiachia G, Marana AN, Ruf T, Ernst A (2011) Census histograms: a simple feature extraction and matching approach for face recognition. Int J Pattern Recognit Artif Intell 25(08):1337–1348
31. Choudhury SK, Sa PK, Padhy RP, Sharma S, Bakshi S (2018) Improved pedestrian detection using motion segmentation and silhouette orientation. Multimed Tools Appl 77(11):13,075–13,114
32. Cotrina C, Bazán K, Oblitas J, Avila-George H, Castro W (2018) Using machine learning techniques and different color spaces for the classification of cape gooseberry (physalis peruviana l.) fruits according to ripeness level. PeerJ PrePrints

33. Cuimei L, Zhiliang Q, Nan J, Jianhua W (2017) Human face detection algorithm via haar cascade classifier combined with three additional classifiers. In: 2017 13th IEEE international conference on electronic measurement & instruments (ICEMI), IEEE, pp 483–487
34. Dai Y, Nakano Y (1996) Face-texture model based on sgld and its application in face detection in a color scene. Pattern Recognit 29(6):1007–1017
35. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE, vol 1, pp 886–893
36. Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. In: European conference on computer vision. Springer, New York, pp 428–441
37. Dargan S, Kumar M, Ayyagari MR, Kumar G (2019) A survey of deep learning and its applications: a new paradigm to machine learning. Archiv Comput Meth Eng pp 1–22
38. Das D, Saharia D, et al. (2014) Implementation and performance evaluation of background subtraction algorithms. arXiv:14051815
39. De Souza F, Pedrini H (2017) Detection of violent events in video sequences based on census transform histogram. In: 2017 30th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), IEEE, pp 323–329
40. Deng Y, Luo P, Loy CC, Tang X (2014) Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM international conference on multimedia, pp 789–792
41. Dollar P, Wojek C, Schiele B, Perona P (2011) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761
42. Dong L, Dong W, Feng N, Mao M, Chen L, Kong G (2017) Color space quantization-based clustering for image retrieval. Front Comput Sci 11(6):1023–1035
43. Dow CR, Ngo HH, Lee LH, Lai PY, Wang KC, Bui VT (2020) A crosswalk pedestrian recognition system by using deep learning and zebra-crossing recognition techniques. Softw Pract Exp 50(5): 630–644
44. Duda RO, Hart PE, Stork DG (2012) Pattern classification. Wiley, New York
45. Dwivedi N, Singh DK, Kushwaha DS (2020) Orientation invariant skeleton feature (oisf): a new feature for human activity recognition. Multimed Tools Appl
46. El-dosuky MA, Oliva D, Hassanien AE (2020) An artificial intelligence system for apple fruit disease classification based on support vector machine and cockroach swarm optimization. In: Joint european-US workshop on applications of invariance in computer vision. Springer, New York, pp 137–147
47. Elgammal A, Muang C, Hu D (2009) Skin detection-a short tutorial. Encycloped Biomet 4:1218–1224
48. Endah SN, Kusumaningrum R, Wibawa HA (2017) Color space to detect skin image: the procedure and implication. Scientif J Inform 4(2):143–149
49. Everingham M, Eslami SA, Van Gool L, Williams CK, Winn J, Zisserman A (2015) The pascal visual object classes challenge: a retrospective. Int J Comput Vision 111(1):98–136
50. Feraund R, Bernier OJ, Viallet JE, Collobert M (2001) A fast and accurate face detector based on neural networks. IEEE Trans Pattern Anal Mach Intell 23(1):42–53
51. Ferryman J, Shahrokni A (2009) Pets2009: dataset and challenge. In: 2009 twelfth IEEE international workshop on performance evaluation of tracking and surveillance, IEEE, pp 1–6
52. Firoze A, Deb T (2018) Face recognition time reduction based on partitioned faces without compromising accuracy and a review of state-of-the-art face recognition approaches. In: Proceedings of the 2018 international conference on image and graphics processing, pp 14–21
53. Freund Y, Schapire RE (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. In: European conference on computational learning theory. Springer, New York, pp 23–37
54. Gaikwad V, Lokhande S (2015) Vision based pedestrian detection for advanced driver assistance. Procedia Comput Sci 46:321–328
55. Gajjar V, Gurnani A, Khandhediya Y (2017) Human detection and tracking for video surveillance: a cognitive science approach. In: Proceedings of the IEEE international conference on computer vision workshops, pp 2805–2809
56. Garcia-Garcia B, Bouwmans T, Silva AJR (2020) Background subtraction in real applications: challenges, current models and future directions. Comput Sci Rev 35:100,204
57. Garcia-Martin A, Martínez JM (2012) On collaborative people detection and tracking in complex scenarios. Image Vision Comput 30(4-5):345–354
58. Ghazali KHB, Ma J, Xiao R, et al. (2012) An innovative face detection based on ycgcr color space. Phys Procedia 25:2116–2124
59. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448

60. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
61. Gomez G, Sanchez M, Sucar LE (2002) On selecting an appropriate colour space for skin detection. In: Mexican international conference on artificial intelligence. Springer, New York, pp 69–78
62. Gonzales RC, Woods RE (2002) Digital image processing
63. González A, Fang Z, Socarras Y, Serrat J, Vázquez D, Xu J, López AM (2016) Pedestrian detection at day/night time with visible and fir cameras: a comparison. Sensors 16(6):820
64. Greche L, Es-Sbai N (2016) Automatic system for facial expression recognition based histogram of oriented gradient and normalized cross correlation. In: 2016 international conference on information technology for organizations development (IT4OD), IEEE, pp 1–5
65. Guo J, Wang J, Bai R, Zhang Y, Li Y (2017) A new moving object detection method based on frame-difference and background subtraction. In: IOP conference series: materials science and engineering, vol 242. IOP Publishing, Bristol, p 012115
66. Haiyuan W, Chen Q, Yachida M (1999) Face detection from color images using a fuzzy pattern matching method. IEEE Trans Pattern Anal Mach Intell 21(6):557–563. https://doi.org/10.1109/34.771326
67. Han H, Tong M (2013) Human detection based on optical flow and spare geometric flow. In: 2013 seventh international conference on image and graphics, IEEE, pp 459–464
68. Han H, Tong M (2013) Human detection based on optical flow and spare geometric flow. In: 2013 seventh international conference on image and graphics, pp 459–464
69. Hapsari G, Prabuwono AS (2010) Human motion recognition in real-time surveillance system: a review. J Appl Sci (Faisalabad) 10(22):2793–2798
70. Hassanpour R, Shahbahrami A, Wong S (2008) Adaptive gaussian mixture model for skin color segmentation. World Academy Sci Eng Technol 41:1–6
71. He K, Gkioxari G, Dollár P, Girshick R (2017a) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
72. He M, Luo H, Chang Z, Hui B (2017b) Pedestrian detection with semantic regions of interest. Sensors 17(11):2699
73. Hoang VD, Hernandez DC, Jo KH (2014) Partially obscured human detection based on component detectors using multiple feature descriptors. In: International conference on intelligent computing. Springer, New York, pp 338–344
74. Horprasert T, Harwood D, Davis LS (1999) A statistical approach for real-time robust background subtraction and shadow detection. In: Ieee iccv, vol 99. Citeseer, New Jersey, pp 1–19
75. Hsu FC, Gubbi J, Palaniswami M (2013) Human head detection using histograms of oriented optical flow in low quality videos with occlusion. In: 2013, 7th international conference on signal processing and communication systems (ICSPCS), IEEE, pp 1–6
76. Jedynak B, Zheng H, Daoudi M (2003) Statistical models for skin detection. In: 2003 conference on computer vision and pattern recognition workshop, IEEE, vol 8, pp 92–92
77. Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. Int J Comput Vis 46(1):81–96
78. Kahu SY, Raut RB, Bhurchandi KM (2019) Review and evaluation of color spaces for image/video compression. Color Res Appl 44(1):8–33
79. Kakumanu P, Makrogiannis S, Bourbakis N (2007) A survey of skin-color modeling and detection methods. Pattern Recognit 40(3):1106–1122
80. Kale K, Pawar S, Dhulekar P (2015) Moving object tracking using optical flow and motion vector estimation. In: 2015 4th international conference on reliability, infocom technologies and optimization (ICRITO)(trends and future directions), IEEE, pp 1–6
81. Kalwa U, Legner C, Kong T, Pandey S (2019) Skin cancer diagnostics with an all-inclusive smartphone application. Symmetry 11(6):790
82. Khalifa AF, Badr E, Elmahdy HN (2019) A survey on human detection surveillance systems for raspberry pi. Image Vis Comput 85:1–13
83. Khan MA, Javed K, Khan SA, Saba T, Habib U, Khan JA, Abbasi AA (2020) Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimed Tools Appl pp 1–27
84. Khan R, Hanbury A, Stöttinger J, Bais A (2012) Color based skin classification. Pattern Recogn Lett 33(2):157–163
85. Khelalef A, Ababsa F, Benoudjit N (2019) An efficient human activity recognition technique based on deep learning. Pattern Recognit Image Anal 29(4):702–715
86. Kim B, Yuvaraj N, Sri Preethaa K, Santhosh R, Sabari A (2020) Enhanced pedestrian detection using optimized deep convolution neural network for smart building surveillance. Soft Comput pp 1–12

87. Kim HK, Park JH, Jung HY (2018) An efficient color space for deep-learning based traffic light recognition. J Adv Transp p 2018
88. Kumar SH, Sivaprakash P (2013) New approach for action recognition using motion based features. In: 2013 IEEE conference on information & communication technologies, IEEE, pp 1247–1252
89. Kwon D, Kim H, Kim J, Suh SC, Kim I, Kim KJ (2019) A survey of deep learning-based network anomaly detection. Clust Comput pp 1–13
90. Ladjailia A, Bouchrika I, Merouani HF, Harrati N, Mahfouf Z (2019) Human activity recognition via optical flow: decomposing activities into basic actions. Neural Comput Applic pp 1–14
91. Lee JY, Yoo SI (2002) An elliptical boundary model for skin color detection. In: Proceedings of the 2002 international conference on imaging science, systems, and technology
92. Li H, Lin K, Bai J, Li A, Yu J (2019) Small object detection algorithm based on feature pyramid-enhanced fusion ssd. Complexity p 2019
93. Li Z, Peng C, Yu G, Zhang X, Deng Y, Sun J (2017) Light-head r-cnn: In defense of two-stage object detector. arXiv:171107264
94. Liu CL, Lee CH, Lin PM (2010) A fall detection system using k-nearest neighbor classifier. Expert Syst Appl 37(10):7174–7181
95. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M (2020) Deep learning for generic object detection: a survey. Int J Comput Vision 128(2):261–318
96. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: single shot multibox detector. In: European conference on computer vision. Springer, New York, pp 21–37
97. Loesdau M, Chabrier S, Gabillon A (2017) Chromatic indices in the normalized rgb color space. In: 2017 International conference on digital image computing: techniques and applications (DICTA), IEEE, pp 1–8
98. Lumini A, Nanni L (2018) Fair comparison of skin detection approaches on publicly available datasets. arXiv:180202531
99. Luo R (2016) Encyclopedia of color science and technology. Springer Publishing Company, Incorporated
100. Luo X, Guan Q, Tan H, Gao L, Wang Z, Luo X (2017) Simultaneous indoor tracking and activity recognition using pyroelectric infrared sensors. Sensors 17(8):1738
101. Ma M (2020) Infrared pedestrian detection algorithm based on multimedia image recombination and matrix restoration. Multimed Tools Appl 79(13):9267–9282
102. Maheswari S, Korah R (2017) Enhanced skin tone detection using heuristic thresholding
103. Mahmoodi MR (2017) Fast and efficient skin detection for facial detection. arXiv:170105595
104. Mateus A, Ribeiro D, Miraldo P, Nascimento JC (2019) Efficient and robust pedestrian detection using deep learning for human-aware navigation. Robot Auton Syst 113:23–37
105. Matsumura R, Hanazawa A (2019) Human detection using color contrast-based histograms of oriented gradients
106. Mehta R, Ozturk C (2018) Object detection at 200 frames per second. In: Proceedings of the european conference on computer vision (ECCV), pp 0–0
107. Mohan A, Papageorgiou C, Poggio T (2001) Example-based object detection in images by components. IEEE Trans Pattern Anal Mach Intell 23(4):349–361
108. Montufar-Chaveznava R (2006) Face tracking using a polling strategy. Proc World Acad Scien Enginner Techn 18:161–165
109. Mu Y, Yan S, Liu Y, Huang T, Zhou B (2008) Discriminative local binary patterns for human detection in personal album. In: 2008 IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8
110. Mulyanto A, Borman RI, Prasetyawan P, Jatmiko W, Mursanto P (2019) Real-time human detection and tracking using two sequential frames for advanced driver assistance system. In: 2019 3rd international conference on informatics and computational sciences (ICICoS), pp 1–5
111. Munder S, Gavrila DM (2006) An experimental study on pedestrian classification. IEEE Trans Pattern Anal Mach Intell 28(11):1863–1868
112. Nambiar A, Taiana M, Figueira D, Nascimento JC, Bernardino A (2014) A multi-camera video dataset for research on high-definition surveillance. Int J Mach Intell Sens Signal Process 1(3):267–286
113. Nguyen DT, Li W, Ogunbona PO (2016) Human detection from images and videos: a survey. Pattern Recogn 51:148–175
114. Ogale NA (2006) A survey of techniques for human detection from video. Survey Univ Maryland 125(133):19
115. Ojha U, Adhikari U, Singh DK (2017) Image annotation using deep learning: A review. In: 2017 international conference on intelligent computing and control (I2C2), IEEE, pp 1–5
116. Ouyang W, Wang X (2012) A discriminative deep model for pedestrian detection with occlusion handling. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp 3258–3265

117. Pang Y, Yuan Y, Li X, Pan J (2011) Efficient hog human detection. Signal Process 91(4):773–781
118. Papageorgiou CP, Oren M, Poggio T (1998) A general framework for object detection. In: Sixth international conference on computer vision (IEEE Cat. No. 98CH36271), IEEE, pp 555–562
119. Paul M, Haque SM, Chakraborty S (2013) Human detection in surveillance videos and its applications-a review. EURASIP J Adv Signal Process 2013(1):176
120. Petrović N, Jovanov L, Pižurica A, Philips W (2008) Object tracking using naive bayesian classifiers. In: International conference on advanced concepts for intelligent vision systems. Springer, New York, pp 775–784
121. Peyré G, Mallat S (2004) Second generation bandelets and their application to image and 3d meshes compression. Math Image Anal MIA p 4
122. Phung SL, Chai D, Bouzerdoum A (2001) A universal and robust human skin color model using neural networks. In: IJCNN'01. international joint conference on neural networks. proceedings (Cat. No. 01CH37222), IEEE, vol 4, pp 2844–2849
123. Phung SL, Bouzerdoum A, Chai D, Watson A (2004) Naive bayes face-nonface classifier: a study of preprocessing and feature extraction techniques. In: 2004 international conference on image processing, 2004. ICIP'04., IEEE, vol 2, pp 1385–1388
124. Piccardi M (2004) Background subtraction techniques: a review. In: 2004 IEEE international conference on systems, man and cybernetics (IEEE cat. no. 04CH37583), IEEE, vol 4, pp 3099–3104
125. Popov A, Dimitrova D (2008) A new approach for finding face features in color images. In: 2008 4th international IEEE conference intelligent systems, IEEE, vol 2, pp 12–33
126. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Shyu ML, Chen SC, Iyengar S (2018) A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surveys (CSUR) 51(5):1–36
127. Rahimzadeganasl A, Sertel E (2017) Automatic building detection based on cie luv color space using very high resolution pleiades images. In: 2017 25th signal processing and communications applications conference (SIU), IEEE, pp 1–4
128. Reddy RVK, Raju KP, Kumar LR, Kumar MJ (2016) Grey level to rgb using ycbcr color space technique. Int J Comput Appl 147(7)
129. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
130. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:180402767
131. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
132. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
133. Riaz I, Piao J, Shin H (2013) Human detection by using centrist features for thermal images. In: International conference computer graphics visualization computer vision and image processing. Citeseer, New Jersey
134. Rusia MK, Singh DK, Ansari MA (2019) Human face identification using lbp and haar-like features for real time attendance monitoring. In: 2019 fifth international conference on image information processing (ICIIP), IEEE, pp 612–616
135. Sayed U, Mofaddel MA, Bakheet S, El-Zohry Z (2018) An elliptical boundary skin model for hand detection based on hsv color space
136. Schmidt A, Kasiński A (2007) The performance of the haar cascade classifiers applied to the face and eyes detection. In: Computer recognition systems, vol 2. Springer, New York, pp 816–823
137. Schneider N, Gavrila DM (2013) Pedestrian path prediction with recursive bayesian filters: a comparative study. In: German conference on pattern recognition. Springer, New York, pp 174–183
138. Schwerdt K, Crowley JL (2000) Robust face tracking using color. In: Proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), IEEE, pp 90–95
139. Sebe N, Cohen I, Huang TS, Gevers T (2004) Skin detection: a bayesian network approach. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004., IEEE, vol 2, pp 903–906
140. Seemanthini K, Manjunath S (2018) Human detection and tracking using hog for action recognition. Procedia Comput Sci 132:1317–1326
141. Senst T, Evangelio RH, Sikora T (2011) Detecting people carrying objects based on an optical flow motion model. In: 2011 IEEE workshop on applications of computer vision (WACV), IEEE, pp 301–306

142. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv:13126229
143. Setjo CH, Achmad B, et al. (2017) Thermal image human detection using haar-cascade classifier. In: 2017 7th international annual engineering seminar (InAES), IEEE, pp 1–6
144. Shaik KB, Ganesan P, Kalist V, Sathish B, Jenitha JMM (2015) Comparative study of skin color detection and segmentation in hsv and ycbcr color space. Procedia Comput Sci 57(12):41–48
145. Sharma SK, Agrawal R, Srivastava S, Singh DK (2017) Review of human detection techniques in night vision. In: 2017 international conference on wireless communications, signal processing and networking (WiSPNET), IEEE, pp 2216–2220
146. Singh DK (2015) Recognizing hand gestures for human computer interaction. In: 2015 international conference on communications and signal processing (ICCSP), IEEE, pp 0379–0382
147. Singh DK (2017) Gaussian elliptical fitting based skin color modeling for human detection. In: 2017 IEEE 8th control and system graduate research colloquium (ICSGRC), IEEE, pp 197–201
148. Singh DK (2018) Human action recognition in video. In: International conference on advanced informatics for computing research. Springer, New York, pp 54–66
149. Singh DK, Kushwaha DS (2016a) Analysis of face feature based human detection techniques. Int J Control Theory Appl 9(22):173–180
150. Singh DK, Kushwaha DS (2016b) Ilut based skin colour modelling for human detection. Indian J Sci Technol 9:32
151. Singh DK, Kushwaha DS (2017) Automatic intruder combat system: a way to smart border surveillance. Def Sci J 67(1):50
152. Sreenu G, Durai MS (2019) Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data 6(1):48
153. Stauffer C, Grimson WEL (1999) Adaptive background mixture models for real-time tracking. In: Proceedings 1999 IEEE computer society conference on computer vision and pattern recognition (cat. no PR00149), IEEE, vol 2, pp 246–252
154. Störring M, Kočka T, Andersen HJ, Granum E (2003) Tracking regions of human skin through illumination changes. Pattern Recognit Lett 24(11):1715–1723
155. Subban R, Mishra R (2013) Combining color spaces for human skin detection in color images using skin cluster classifier. In: International conference on advances in recent technologies in electrical and electronics. Citeseer, New Jersey, pp 68–73
156. Sultana F, Sufian A, Dutta P (2020) A review of object detection models based on convolutional neural network. In: Intelligent computing: image processing based applications. Springer, New York, pp 1–16
157. Sun D, Roth S, Lewis J, Black MJ (2008) Learning optical flow. In: European conference on computer vision. Springer, New York, pp 83–97
158. Tamgade SN, Bora VR (2009) Motion vector estimation of video image by pyramidal implementation of lucas kanade optical flow. In: 2009 second international conference on emerging trends in engineering & technology, IEEE, pp 914–917
159. Tang J, Deng C, Huang GB (2015) Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst 27(4):809–821
160. Tang S, Goto S (2009) Human detection using motion and appearance based feature. In: 2009 7th international conference on information, communications and signal processing, (ICICS), IEEE, pp 1–4
161. Teixeira T, Dublon G, Savvides A (2010) A survey of human-sensing: methods for detecting presence, count, location, track, and identity. ACM Comput Surv 5(1):59–69
162. Terrillon JC, David M, Akamatsu S (1998) Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments. In: Proceedings third IEEE international conference on automatic face and gesture recognition, IEEE, pp 112–117
163. Terrillon JC, Shirazi MN, Fukamachi H, Akamatsu S (2000) Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In: Proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), IEEE, pp 54–61
164. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vision 104(2):154–171
165. Vezhnevets V, Sazonov V, Andreeva A (2003) A survey on pixel-based skin color detection techniques. In: Proceedings graphicon, Moscow, Russia, vol 3, pp 85–92
166. Viola P, Jones M, et al. (2001) Robust real-time object detection. Int J Comput Vision 4(34-47):4
167. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: a brief review. Comput Intell Neurosci p 2018
168. Walia GS, Kapoor R (2014) Human detection in video and images—a state-of-the-art survey. Int J Pattern Recognit Artif Intell 28(03):1455,004

169. Wang L, Shi J, Song G, Shen IF (2007) Object detection combining recognition and segmentation. In: Asian conference on computer vision. Springer, New York, pp 189–199
170. Wang X, Han TX, Yan S (2009) An hog-lbp human detector with partial occlusion handling. In: 2009 IEEE 12th international conference on computer vision, IEEE, pp 32–39
171. Wang X, Shen C, Li H, Xu S (2019) Human detection aided by deeply learned semantic masks. IEEE Trans Circ Syst Video Technol
172. Wang Y, Yuan B (2001) A novel approach for human face detection from color images under complex background. Pattern Recogn 34(10):1983–1992
173. Womg A, Shafiee MJ, Li F, Chwyl B (2018) Tiny ssd: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection. In: 2018 15th conference on computer and robot vision (CRV), IEEE, pp 95–101
174. Wong WK, Hui JH, Loo CK, Lim WS (2011) Off-time swimming pool surveillance using thermal imaging system. In: 2011 IEEE international conference on signal and image processing applications (ICSIPA) IEEE, pp 366–371
175. Wong WK, Hui JH, Desa JBM, Ishak NINB, Sulaiman AB, Nor YBM (2012) Face detection in thermal imaging using head curve geometry. In: 2012 5th international congress on image and signal processing, IEEE, pp 881–884
176. Wu B, Nevatia R (2005) Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: Tenth IEEE international conference on computer vision (ICCV'05) volume 1, IEEE, vol 1, pp 90–97
177. Wu B, Nevatia R (2007) Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int J Comput Vis 75(2):247–266
178. Wu J, Geyer C, Rehg JM (2011) Real-time human detection using contour cues. In: 2011 IEEE international conference on robotics and automation, IEEE, pp 860–867
179. Wu X, Sahoo D, Hoi SC (2020) Recent advances in deep learning for object detection. Neurocomputing
180. Yamashita A, Ito Y, Kaneko T, Asama H (2011) Human tracking with multiple cameras based on face detection and mean shift. In: 2011 IEEE international conference on robotics and biomimetics, IEEE, pp 1664–1671
181. Yang B, Lei Y, Yan B (2015) Distributed multi-human location algorithm using naive bayes classifier for a binary pyroelectric infrared sensor tracking system. IEEE Sensors J 16(1):216–223
182. Yang J, Lu W, Waibel A (1998) Skin-color modeling and adaptation. In: Asian conference on computer vision. Springer, New York, pp 687–694
183. Yang MH, Ahuja N (1998) Gaussian mixture model for human skin color and its applications in image and video databases. In: Storage and retrieval for image and video databases VII, international society for optics and photonics, vol 3656, pp 458–466
184. Yuan B, Li S (2017) Extended census transform histogram for land-use scene classification. J Appl Remote Sens 11(2):025,003
185. Zabih R, Woodfill J (1994) Non-parametric local transforms for computing visual correspondence. In: European conference on computer vision. Springer, New York, pp 151–158
186. Zhang B, Horváth I, Molenbroek JF, Snijders C (2010) Using artificial neural networks for human body posture prediction. Int J Ind Ergon 40(4):414–424
187. Zhang H, Hong X (2019) Recent progresses on object detection: a brief review. Multimed Tools Appl 78(19):27,809–27,847
188. Zhang H, Zhao D, Gao W, Chen X (2000) Combining skin color model and neural network for rotation invariant face detection. In: International conference on multimodal interfaces. Springer, New York, pp 237–244
189. Zheng J, Ranjan R, Chen CH, Chen JC, Castillo CD, Chellappa R (2020) An automatic system for unconstrained video-based face recognition. IEEE Trans Biomet Behav Ident Sci 2(3):194–209
190. Zhu A, Wang T, Qiao T (2019) Multiple human upper bodies detection via candidate-region convolutional neural network. Multimed Tools Appl 78(12):16,077–16,096