INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.147 Volume 15, 2021

# Segmentation and selective feature extraction for human detection to the direction of action recognition

Lakhyadeep Konwar[1], Anjan Kumar Talukdar[2], Kandarpa Kumar Sarma[3], Navajit Saikia[4], Subhash Chandra Rajbangshi[5]
[1, 2, 3, 5] Dept. of ECE, GUIST, Gauhati University, Jalukbari, Assam, India, 781014
[4] Dept. of ETE, Assam Engineering College, Jalukbari, Assam, India, 781013

Received: February 15, 2021. Revised: August 25, 2021. Accepted: September 6, 2021. Published: September 8, 2021.

**Abstract- Detection as well as classification of different object for machine vision application is a challenging task. Similar to the other object detection and classification task, human detection concept provides a major role for the advancement in the design of an automatic visual surveillance system (AVSS). For the future automation system if it is possible to include human detection and tracking, human action recognition, usual as well as unusual event recognition etc. concept for future AVSS, it will be a greater success in the transformable world. In this paper we have proposed a proper human detection and tracking technique for human action recognition toward the design of AVSS. Here we use median filter for noise removal, graph cut for segment the human images, mathematical morphology to refine the segmentation mask, extract selective feature points by sing HOG, classify human objects by using SVM with polynomial kernel and finally particle filter for tracking those of detected human. Due to the above mentioned combinations our system can independent to the variations of lightening conditions, color, shape, size, clothing etc. and can handle the occlusion. Our system can easily detect and track human in different indoor as well as outdoor environment with a automatic multiple human detection rate of 97.61% and total multiple human detection and tracking accuracy is about 92% for AVSS. Due to the use of HOG to extract features after graph cut segmentation operation, our system requires less memory for store the trained data therefore processing speed as well as accuracy of detection and tracking will be better than other techniques which can be suitable for action classification task.**

**Keywords- Action recognition, Human detection, Occlusion handling, Segmentation.**

## I. Introduction

In today's world crimes and terrorist activities are increases in an uncontrolled manner. Normal people are unable to live freely in any society as they want. Common people are afraid to live freely inside as well as outside home because now a day's various kind of unusual activities such as rape, murder, kidnapping, fighting, arms trafficking, robbery are happening in various crowded places such as shopping mall, home, highways, daily market, parking places etc. Now a day's various object detection as well as classification is possible trought various machine learning as well as deep learning algorithm. Detection of object is a challenging task for machine vision applications. It is difficult to detect an object from an image if there is a crowded scene in the image [8]. For that purpose we have to use properly trained classifier, so that proper detection of that kind of object easily takes place [10]. Human detection from image is a newly identified challenging task owing to their different poses, color, size, shape [12] etc. It is also difficult to detect human from images due to the variable appearance [15] and disappearance of human in different images. Some other technique that has already been implemented gives some amount of accuracy but system provides slow response. Some other techniques are there where the system is unable to work in various lightening conditions as well as outdoor environment [19]. So we have to implement algorithm where the accuracy will be better, system provides better response, provides proper result of detection which is independent to the lightening changes, works properly in indoor as well as outdoor environments and properly works in crowded scenes. We have to also focus that the system should be cheaper as well as used friendly. Through the use of those machine learning and deep learning algorithm it is also possible to classify human actions after the detection stage of human from video footage. Here the system can be further designed for real time usual and unusual action classification for Automatic Visual Surveillance System (AVSS) design.

Lot's of resesrcher and scientist are working in the area of designing a proper AVSS that includes human detection [12], [14], usual and unusual event detection [21], behaviour recognition [22], [43] , action recognition or activity recognition [23], [24], human tracking [5], [20], [25], [26], [28] etc.

Ahmad et al. [3] proposed a method for segment detected moving object from image sequences using mean

shift clustering that was mainly applicable to the detection of motion changes for those of the detected objects. Due to the used of Mean shift clustering based segmentation technique by Ubukata et al. [4] for segment human object for detection purpose the human detection accuracy as well as occlusion handle capacity is increased. The background subtraction and foreground segmentation method for object detection which is based on the Gaussian Mixture Model (GMM) is introduced by Thombre et al. [5] and Hafiz et al. [6] that can applicable for shadow removal by using the contrast adjustment technique. Bokov et al. [7] used graph cut segmentation techniques to segment boundary and regions of an image from an N-dimensional images. In [8] Zhang used graph cut segmentation technique to segment multiple objects that can be used for occlusion handling in the area of tracking of multiple moving object from videos. Combination of graph cut based object segmentation and Histogram of Oriented Gradient (HOG) for proper human detection was introduced by Lakshmi et al. [9]. After that Ramya et al. [10] and Kharabe et al. [11] found that graph cut segmentation technique provides proper result for human as well as detection of other objects in the case of moving object detection purpose. The first concept of human detection was introduced by Dalal et al. [12] that is the combination of HOG feature descriptor and Support Vector Machine (SVM) [13] classifier for proper human detection. They found that HOG feature descriptor provides better result of feature extraction and therefore human detection accuracy is increases. After that a method was introduced by Zhu et al. [15] which is based on variable block size HOG feature descriptor that captures salient features of human object automatically. They identify the appropriate set of blocks from a large set of possible blocks the use of Adabost based feature selection by using integral image representation and a rejection of cascade that significantly speed up the computation. Kachouane et al. [16] presented an algorithm for human detection and recognition in real time images that is based on the combination of HOG feature descriptor and SVM classifier are provides good results of detection and generally used for robotic tasks. A speed up method of pedestrian detection which was based on the two stage cascade structure, i.e combination of HOG and LBF (Local Binary Fitting) was introduced by Park et al. [17]. Here the first stage extracts the features from the regions which are characterized by pedestrian only and therefore the systems pedestrian detection accuracy is three times faster than other conventional techniques. Another system that can properly work in terms of variation of size, potential occlusion, amount of context, noise and clutter situations was developed by Bell et al. [18], where the method can automatically detect dismounted human at long range from a single, highly compressed images. Zhao et al. [19] proposed a method to track multiple human in complex situations by the use of a single stationary video camera where the human motion is decomposed into a global motion (i.e. position and orientation) and limb motion (i.e. more detailed body postures). Here first objective was to segment multiple human objects and track their global motion in complex situations where they may move in small groups, have inter-occlusions, cast shadow on the ground, the reflection etc. and then the second objective is to estimate the locomotion modes (i.e. walking, running, standing etc.) and a $3D$ body postures. At [20] Kushwaha introduced a new algorithm which is based on the combination of Haar feature descriptor for feature extraction and particle filter for tracking. Beyond these combinations they also used binary adaptive boosting for object classification and therefore the system can detect and track multiple human objects in videos that are adequately fast in the presence of variation of poses, shapes, sizes, clothings etc. Beaugendre et al. [25] and Saboune et al. [26] used particle filter based tracking of human for their system implementations. The particle filter based object tracking system provides good estimation of the $3D$ positions with the help of optimization of particles which is depend upon those of the likelihood function applied and therefore this technique can track people that are newly enter to a scene and those are recovering from occlusions. An another technique for multiple object tracking was introduced by Li et al. [27] which is based on the kalman filtering technique. Due to the use of this technique they reduces the search scope and search time of moving objects to achieve the fast tracking. They also established the corresponding relationship through moving object features matching to deal with separation of objects after object merged.

Lao et al. [30] proposed a flexible framework for sematic analysis of human behaviour from a monocular surveillance video. In their system the trajectory estimation and human body modeling technique meets the requirements for the analysis of human activity and events in video sequences. In their paper they also introduced a 3-D reconstruction technique for scene understanding so that the human action can analyzed from different views. Seemanthi et al. [31] introduced a new approach for human object detection based on clustering based segmentation technique. Their system can detect and track human by using HOG and SVM algorithm. Angelini et al. [32] proposed a framework for single as well as multi view action recognition based on the space time volume (STV) of human silhouettes and 3D-HOG. They used PCA over local features as L2 regularized logistic regression (L2-RLR) for learning actions from local features. Thurau et al. [33] proposed a technique for human detection and simultaneous behaviour recognition from images as well as image sequences. Here the author apply clustering algorithm to sequences of HOG of human motion images for action representations. Here they classify the human behaviour based on KL divergence of behaviour histogram. To speed up the action recognition model bag of histogram of optical flow (BoHOF) was propose by Sahoo et al. [34]. Here optical flow is calculated over segmented images. Features are thresholded and bagged to compute the BoHOF. Sobel edge filter was used to remove the shsdow effect, median

filter to supress background noise, extract HOG features from 3D projected planes and combined with the Bo-HOF. Finally SVM with RBF kernel used for classification of different actions. Jagadeesh et al. [35] proposed a method for human detection, tracking, recognition and classifications of actions. Here human detection was performed by GMM, optical flow algorithm for tracking and finally SVM for classification of actions. A novel feature descriptor was introduces by Sargano et al. [36] for multiview human action recognition where the proposed descriptor employes the region based features those are extracted from human silhouette. They used multiclass SVM to classify different actions.

In this paper we have implemented an automatic multiple human detection and tracking algorithm for human action classification to the directions of a proper automatic visual surveillance system (AVSS) design which is independent to the lightening conditions as well as shape variations, handle the occlusion, and improve the human detection and tracking accuracy for human action recognition with less ammount of feature vectors. Our system design involves mainly following parts: (a). data collection, (b). application of median filter for noise removal, (c). maximum flow based graph cut technique for segmentation of human image, (d). mathematical morphology to refine the segmentation mask, (e). HOG as a feature descriptor, (f). SVM as a linear classification purpose and (g). finally particle filter based technique for human tracking. Due to the combinations of these, our system provides better accuracy as well as system provides better performance (less time requirements). This system is also robustly familiar with variation of lightening condition based detection, handle occlusions as well as it works in any environments that can be used for human action recognition.

The remaining part of this paper is systemized in the following ways: Section II provides the step by step theoretical considerations of proposed work on multiple human detection and tracking to the direction of action classifications for AVSS design, section III provides the experimental setups required for the system design, section IV provides the experimental results and analysis of proposed system, section V shows performance evaluations of the designed system and finally section VI concludes the paper.

## II. Theoretical Considerations

Our main aim is to design a proper human detection system for activity/ action recognition purpose to the directions of automatic visual surveillance system design. Therefore for proper system design we have to consider some step by step methods. Figure 1 shows the pipeline of human detection and tracking for action recognition model; where first one is the input video dataset section, than next one is the feature extraction section, third one is the classification section and fourth one is the tracking section. Figure 2 shows the pipeline of training system that comprised of input dataset section (positive as well as negative), segmentation section, feature extraction section and SVM classifier training section.

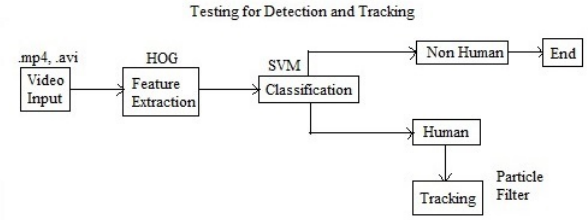


Fig. 1: Pipeline of human detection and tracking for action recognition model

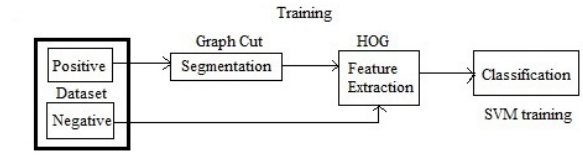


Fig. 2: Pipeline of the training system

In the training part we collected some positive as well as negative dataset separately and than we apply max cut based graph cut segmentation to segment the human images from positive dataset and than apply mathematical morphological operations opening, closing, erosion and dilation to reconstruct the loss portion of images. After that we extract some strong features from all of the datasets (positives as well as negative) by using HOG feature descriptor. Then those of the extracted features are trained by a linear SVM classifier and store it to our system. In the testing part we test human are exist or not in real time videos that are captured by a mobile phone camera. Finally we can track those of the detected human by using the particle filter. So that robust automatic multiple human detection and tracking for action clasiification can takes place which can be further used for automatic visual surveillance system. The major blocks involves in our proposed system are discussed below:

### *A. Noise removal*

When an image is captured by a camera, there are lots of images are present(i.e. gaussian noise, salt and paper noise, speckle noise, poisson noise etc.). In general linear filters are used for noise reduction as well as noise removal process. Non-linear filters can also be used for noise removal purpose. We use median filter for our system. In general median filter is a non-linear digital filtering technique that can be often used for such type of noise (as mentiuoned above) removal purpose. The main background of median filter is to run through the signal entri by entry, by replacing each of the entry with the median of neighbouring entries [1]. Median filter has some advantages over linear techniques that it can eliminate the input noise with extremely large magnitudes.

### B. Graph Cut Segmentation

Graph cut segmentation is one of the image segmentation technique can that properly extract foreground from an image by eliminating the background [9]. Graph cut technique is basically based on the graph theory that can minimizes the energy function by the use of max-flow min-cut theorem [10]. A graph is a set of vertices $V$ and edge $E$ that connect various pairs of vertices. A graph can be written as

$$G = (V, E) \tag{1}$$

Where each edge can be represented by a pair of vertices. i.e

$$E \subset V \times V \tag{2}$$

Here the graphs are often drawn as a set of points with curves connecting the points and the degree of a vertex is the number of edges incident on that vertex.

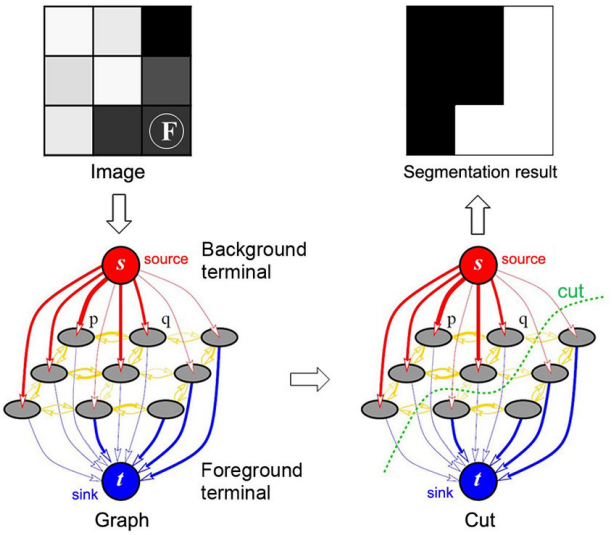


Fig. 3: Graph cut segmentation process for a $3 \times 3$ image [45]

Here each joint of pixel connecting point has a weight from source to destination. The graph cut segmentation process search for the smallest possible weighted path and select it. Figure 3 shows the graph cut segmentation process, where the dotted green line shows the cutted path.

### C. Morphological Operation

We use mathematical morphological operation to refine the segmentation mask by using differently shaped structured elements [2]. The structuring element are a small sets or subimages which is used to probe an image under the study for properties of interest. The structuring elements are asymmetric about the direction of its origin. We use mathematical morphology operations opening, closing, erosion and dilation to reconstruct the loss portions of an image.

### D. Feature Selection

For proper detection of human in an image it is necessary to use a proper feature descriptor. We use the HOG descriptor for our feature extraction process that gives proper result in case of human detection in different lightening conditions or poor lightening conditions or variations of lightening conditions [12]. HOG is one type of feature descriptor which is mostly used to extract some strong feature point depending upon some strong intensity variation points in an image [15]. The basic

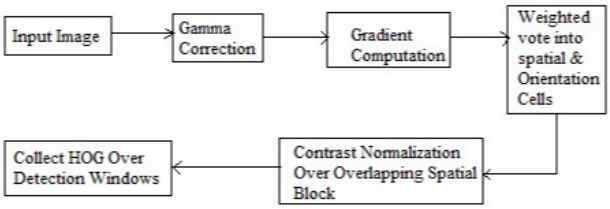


Fig. 4: Basic blocks of HOG feature descriptor

blocks of HOG feature descriptors are shown in figure 3; where there are five major blocks are present. They are: gamma correction, gradient computation, binning orientation, contrast normalization over the overlapping blocks, feature vector collection [12]. Figure 5 shows the graphical representation of HOG feature descriptor application for human detection. Here the detector window is tiled with a grid of overlapping blocks. Each block contains a grid of spatial cells. For each cell, the weighted vote of image gradients in orientation histograms is performed. These are locally normalised and collected in one big feature vector.

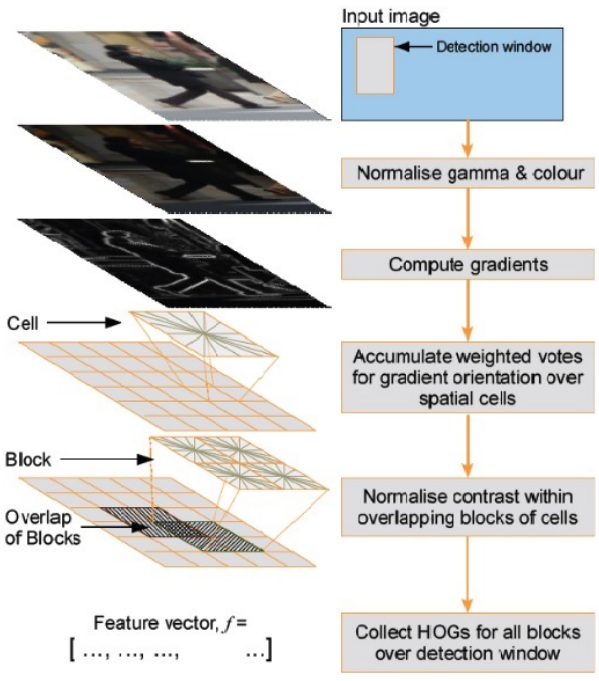


Fig. 5: Graphical representation of each steps of HOG feature descriptor [12]

1) Gamma Correction or Color Normalization
Gamma correction is a non-linear operation which is used to encode and decode luminance or tri-stimulus values in videos that can be defined by the following power-law expression:

$$V_{out} = V_{in}^{\gamma} \tag{3}$$

where $V_{in}$ is the non-negative real input value which is raised to the power $\gamma$ and than multiplied by the constant $A$, to get the output value $V_{out}$. In case of $A = 1$, inputs and outputs are typically in the range between 0 and 1.
The power law equation $V_{out} = V_{in}^{\gamma}$ in, the curve on the logarithmic plot or log-log plot is a straight line where slope is represented by the derivative operator as

$$\gamma = \frac{d\ log(V_{out})}{d\ log(V_{in})} \tag{4}$$

2) Gradient Computations
An image gradient computation is nothing but the directional change in the intensity or color of an image [15]. Mathematically, the gradient of an image intensity function at each of the image point is a 2D vector with some components that can be given by the derivatives in the horizontal and vertical directions [18]. At each stage of the image point in the direction of largest possible change of intensity or increasing of intensity and the length of the feature vector is corresponding to the rate of change in that directions. The gradient image can be generated from original image by the use of a filter convolving process. In case of HOG feature descriptor sobel operator is used. Sobel operator is generally used particularity with the edge detection algorithms where it creates and image emphasis edge. At each of the points of an image, the sobel operator result is either the corresponding gradient vector on the norms of the vector [17]. If $A$ be the source image and $G_x$ and $G_y$ be two images where each point contains the horizontal and vertical derivative approximations. Then we have

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A \tag{5}$$

and

$$G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A \tag{6}$$

Where $*$ denoted the two dimensional $(2D)$ signal processing convolution operation. Than we have the gradient magnitude be

$$G = \sqrt{G_x^2 + G_y^2} \tag{7}$$

and the gradient direction becomes

$$\theta = arctan\left(\frac{G_y}{G_x}\right) \tag{8}$$

Here, $\theta$ is 0 for vertical edge which is lighter on the right side.

3) Orientation Binning
Here the main aim is to calculate the cell histograms where each pixel within the cells that casts a weighted vote for an orientation based histogram channel which is based on the values that is found in the gradient computation. The cell themselves may be either rectangular or radical in shape and the histogram channels are widely spread over 0 to 180° for unsigned gradient that are used with contribution of a histogram channels that performed best for a human detection system. For the weighted vote, pixel contribution can be either the magnitude of the gradient itself or some of the functions of the magnitudes [18].

4) Block Description
The gradient strength must be locally normalized in account of the changes in illumination and the contrast [17][29]. For that purpose grouping of the cell is required which can be groups to larger and are specially connected blocks. These blocks that are used are typically overlap that means each cell of those blocks contributes more than once to the final descriptor. There are two main block of geometrics exist; they are Rectangular HOG (R-HOG) and circular HOG (C-HOG). In general R-HOG blocks are square that can be represented by three parameters and can be describe as the number of cell per blocks, the number of pixels per cell and the number of channels per cell histogram. The R-HOG blocks are generally used in conjuction to encode the spatial form of information that can be mainly use for changing of pixels position or pixels position changing purpose. C-HOG blocks are to be found in two variants, one is with a single, central cell another is those which are with a angularly divided central cell. The C-HOG blocks are to be best described with four parameters; the number of angular and radial bins used, the radius of the centre bin, and the expansion factor that are used for the radius of additional radial bins.

5) Block normalization over overlapping blocks
In general there are four different methods available/ used for block normalization [12]. We use only the L2-norm, where we have to consider a non-normalized vector $v$ that contain all of the histograms in a given block $\|v\|_k$ be its $k^{th}$ number of norms, where $k = 1, 2$ and $e$ be a small constant. Then the normalization factor be the following forms:

$$f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \tag{9}$$

According to Dalal and Trig [12] $L2 - hys$, $L2 - norm$, and $L1 - sqrt$ schemes provides similar performance, while the$L1 - norm$ provides quitely less reliable performance. All of the above mentioned methods provide similar performance over

non-normalized data.

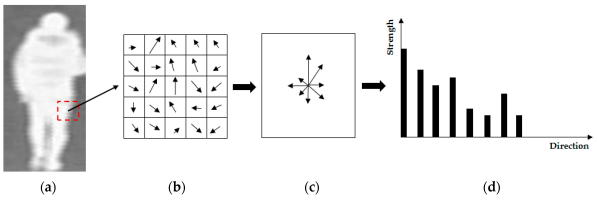


Fig. 6: View of histogram for a small portion of a human image (a) the input image (b)gradient map with gradient strength and direction of a sub block of the input image (c) accumulated gradient orientation and (d) histogram of oriented gradients [41].

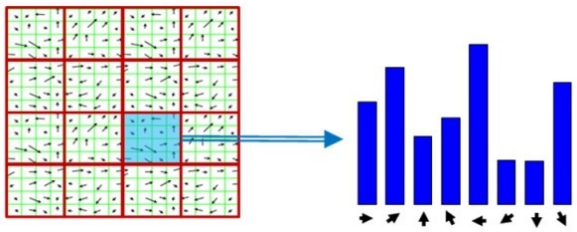

Fig. 7: Histogram of gradient orientation over spatial cells [42]

Figure 6 shows the extraction of feature vector in an small portion of a human image. Here first gradient direction of bins are calculated for each blocks, than orientation of gradient direction is takes place and finally histogram are generated. Figure 7 shows the cell histogram for each bins of each of the cells in a blocks. For each of the cell there will be a histogram.

6) Feature vectors collection
Feature vectors are nothing but the collections of some feature points that are in numerical format. On the other way we can say that feature vectors are some feature values those are in mathematical form.

### E. Classifier Selection

Classifiers are some mathematical algorithms that takes features set as input and produces a class level outputs. SVM is used to classify objects over a hyperplane or by the use of a hyperplane, with the help of an appropriate non-linear mapping $\psi(\cdot)$ for a sufficiently high dimension data that are from different categories can always be separated by a hyperplane [13].

### F. Particle Filter for Tracking

In general Particle filter [20] [25][26][43] uses a generic motion selection sampling with a set of particles, sometimes called samples are used to represent the posterior distribution of some stocastic process that gives noisy and/or partial observations.

## III. Experimental Setup

### A. Hardware System Requirement

Our system is mainly design for video surveillance purpose. Therefore we need some of the basic hardware systems such as single or multiple CCTV (closed circuit television), a control centre, a server designed for human detection and tracking, Ethernet cable and a VDU (visual display unit). Figure 8 shows the basic hardware schematic system that can be used for our system if we will design it for online system. Where a camera is connected to main system (PC of control room)via Ethernet cable. We can use various no. of camera depending upon the requirements but never cross the limitation of connections. We test our system for offline system but our direction is to design an online system. Here the video frames are captured by using a Samsung Galaxy J5 prime mobile phone camera with 25 fps (frames per second) and the resolution of $1024 \times 768$ with colorspace RGB (Red Green Blue). Our proposed system is operating on a Personal computer with Intel Core i3 2350M CPU 2.30 GHz Processor with 4GB RAM with pre-installed Windows 7 Ultimate operating system.

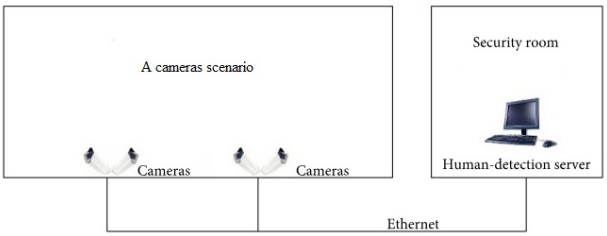


Fig. 8: Basic schematic arrangement of an AVSS system

### B. Software System Requirements

The system is implemented by using OpenCV 3.4.1 library and CodeBlock 10.12 library on C/C++ programming language by using Microsoft Visual Studio 2008. Also we use MATLAB 2017a for some graphical representations.

### C. Dataset preparation

Here all of the positive as well as negative dataset are of size $128 \times 64$. For the dataset section we have arranged as following form:

1) Positive Training Dataset
We use a well established dataset namely MIT pedestrian dataset as a positive dataset which contains 888 human images. Here in these images human are in front views, back view as well as side views of walking style in different environments. Figure 9 shows some sample of positive datasets.

2) Negative Training Dataset
We collect some image dataset from different location over internet. Here a total of 1000 image samples that contains house, car, bike, tree etc. as shown in figure 10.

Fig. 9: Samples of MIT padestrian dataset


Fig. 10: Samples of negative dataset that we have collected

3) Testing Dataset
Here we collect some video captured by a mobile phone camera where the length of each of the video streams are about 5 minutes and all are in .mp4 format. The videos are collected in different environments (corridoor of a medical and research centre, footpath of a highway, campus of an university, on a trunk road etc.). Here a total of 10 video stream are of size $1024 \times 768$ with a frame rate of $25 fps$.

## IV. Results and analysis of proposed system

### A. Segmentation and application of morphological operation

We have apply graph cut segmentation over annotated positive images to segment the selective portion of the human image after the noise removal process using median filter. Figure 11 shows the some samples of annotated human images which is indicated with a yellow color. Here mathematical morphological operation opening, closing, errosion and dilation are applied to reconstruct the loss portions of an image. Figure 12 shows some samples of segmented image and figure 13 shows the sample of binary image. Here segmentation is perform for only the positive training annotated images.

### B. Feature Extraction

Here the feature extraction for the videos are same as training process but here difference is that we have to convert the video to frame first and then extract features by using HOG. Here noise removal part is also done after video to frame conversion process. Table 1 shows some specifications of our HOG feature descriptor that we have use. By using those specifications we get $888 \times 3780$ no. of feature vectors without segmentation and $888 \times 3780$ no. of feature vectors for different 888 positive images after segmentation. But in the feature map for the segmented image comprises of 0's if there is no human images. We also get $1000 \times 3780$ no. of of feature vector for 1000 negative images. From that we have found that we got 3780 no. of feature vector for each images of size

Fig. 11: Samples of some annotated images

Fig. 12: Samples of some segmented images

Fig. 13: Samples of binary images

$128 \times 64$ without segmentation. We get different no. of feature vector for different testing video samples. Here the no. of feature vectors for testing section are huge amount.

Table 1: Some specifications of our HOG feature descriptor

| | |
|---|---|
| Bin size used | 16*16 |
| Bin range | 0 to 255 |
| Color space | RGB |
| Pixel per cell | 16 |
| Cell per block | 4 |
| No of histogram bin | 9 |
| Angle of notation | 0 to 180 degree |
| Histogram range | 0 to 255 |

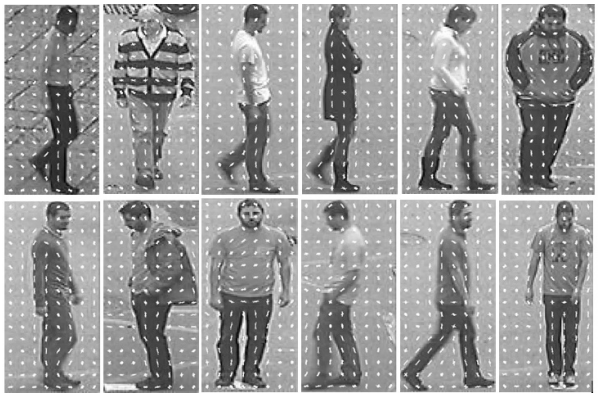

Fig. 14: Position of gradient changes of feature vector over some training positive images without segmentation

Table 1 shows some specifications of HOG feature descriptor that we have used. Figure 14 and 16 shows the position of gradient changes of feature vector of human images as well as non human images without segmentation and figure 15 shows the position of gradient of feature vector of human images after segmentation are calculated by using HOG feature descriptor. Here we have seen that there are some sudden changes of gradient direction which is shown in the positions where sudden intensity changes are occurred. Figure 17 shows the zoomed version of the distribution of gradient changes of feature vector of segmented human images or positive images which is used to train our detector, where the histogram bins are distributed from 0 to 180°. For sudden intensity change there will be a sudden gradient change, therefore we have some feature vector. Figure 18 and 19 shows the samples of feature vector plot for positive after segmentation and negative images. It shows that the values of each feature vector in $Y$ directions against no. of feature vectors in $X$ directions. It shows that 3780 no of feature vectors for positive with segmentation as well as negative images without segmentation, which is distributed from 0 upto 0.78459 for positives and from 0.001231 to 0.471561 for negative images in the case of training images. We got feature values, which are randomly distributed from 0 to 0.4859687 for test images that are converted frame from captured videos as shown in figure 21. Figure 20 shows the feature vector distribution plot for a testing image from a video file.

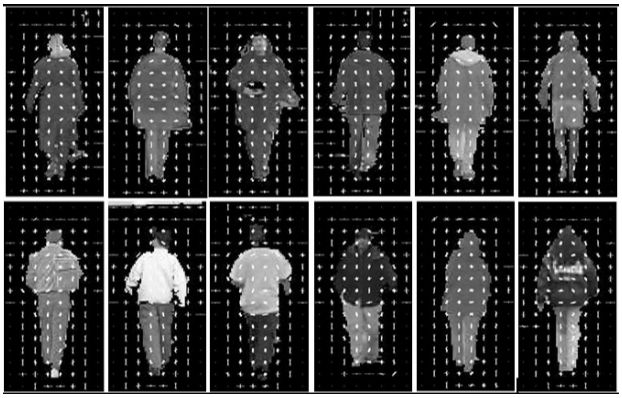

Fig. 15: Position of gradient changes of feature vector over some training positive images after segmentation

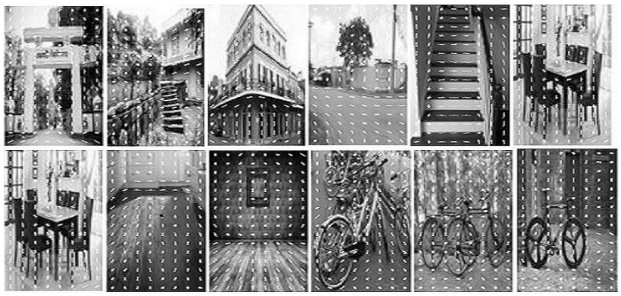

Fig. 16: Position of gradient changes of feature vector over some training positive images

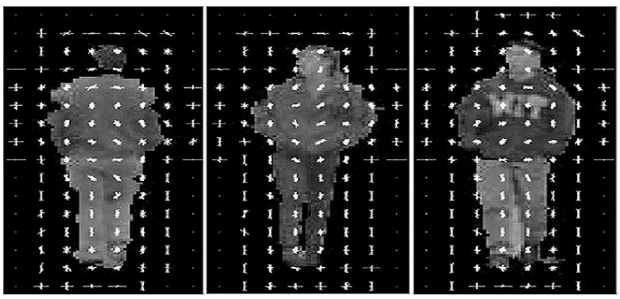

Fig. 17: Magnified position of gradient changes of feature vector over some samples of positive images with segmentation

### *C. Linear SVM based Classification*

Our main goal is to detect and track human in videos or sequence of images for action recognition purpose, therefore we have to classify human and non-human data from our input video dataset. First we classify the human and non human data by using SVM classifier and then pass it to the next step for tracking purpose. Table 2 shows some specifications for our classifier. Here we use linear SVM with polynomial kernel, and human and non-human classes are considered as $+1$ and $-1$. Table

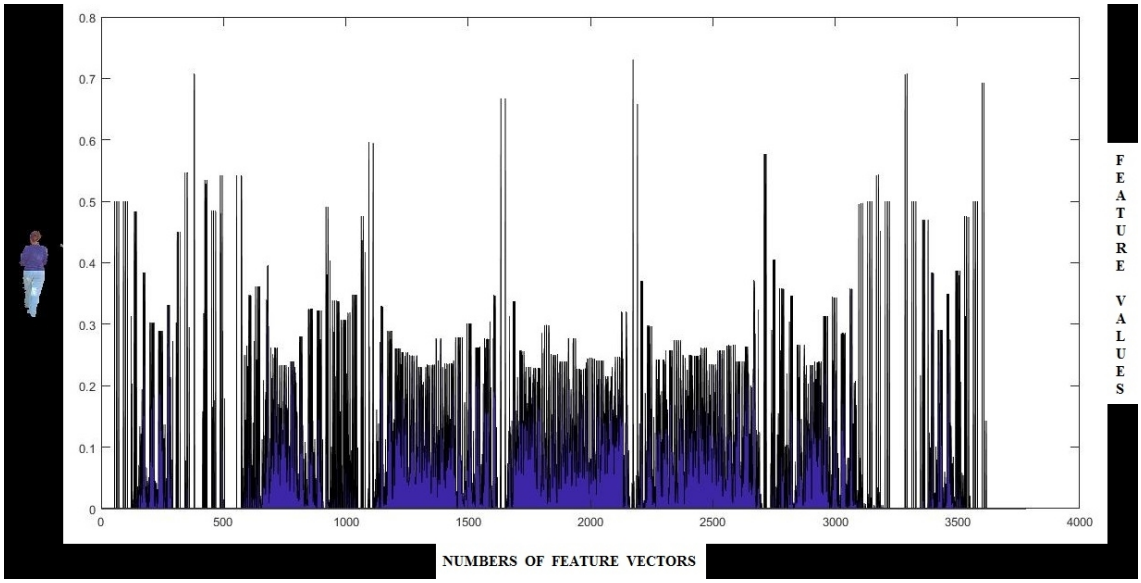


Fig. 18: Feature vector distribution plot of segmented positive images against no. of features

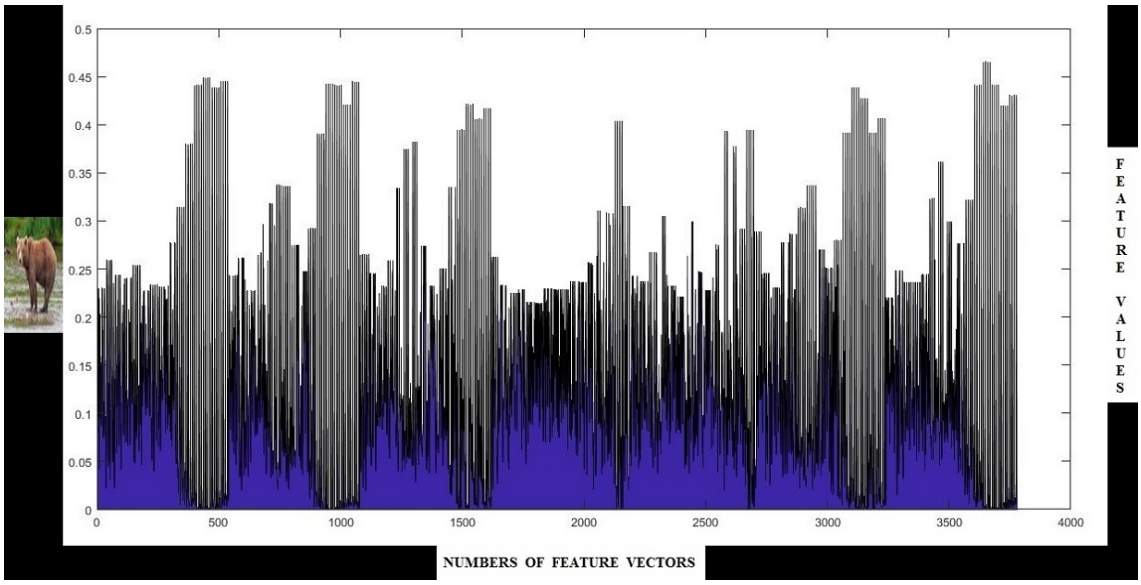


Fig. 19: Feature vector distribution plot of negative images against no. of features

2 shows the specifications that is used to train our SVM classifier.

Table 2: Specifications for our SVM classifier

| | |
|---|---|
| SVM type use | Linear |
| Kernel use | Polynomial |
| Human class | +1 |
| Non-human class | -1 |

### *D. Particle Filter Based Tracking*

The basic algorithm which is used for person tracking are discussed below:

1) Initialization:- First we represent $P(X_0)$ by a set of $N$ number of samples $(s_0^{k,-},\ w_0^{k,-})$, where $s_0^{k,-} \sim P_s(S)$ and $w_0^{k,-} = P(w_0^{k,-})/P_s(S = s_0^{k,-})$. Ideally $P(X_0)$ has a simple form where $s_0^{k,-} \sim P(X_0)$ and $w_0^{k,-} = 1$.
2) Prediction:- Represent $P(X_i \setminus y_0, y_{i-1})$ by using $(s_i^{k,-},\ w_i^{k,-})$. Where $s_i^{k,-} = f(s_{i-1}^{k,+}) + \xi_i^k$ and $\xi_i^k \sim N(0, \sum d_i)$.
3) Correction:- Represent $P(X_i \setminus y_0, y_i)$ by using $(s_i^{k,+},\ w_i^{k,+})$. Where $s_i^{k,+} = s_i^{k,-}$ and $w_i^{k,+} = P(Y_i = y_i \setminus X_i = s_i^{k,-}) w_i^{k,-}$.
4) Re-sampling:- Normalize the weight to set $\sum_i w_i^{k,+} = 1$ then compute the variance of the normalized weights. If the variance exceeds some threshold value then construct a new set of samples by drawing and replacement $N$ numbers of samples from the old set by using the weights as the probability that of a sample will be drown. Then the weight of each samples will be $\frac{1}{N}$.

Fig. 20: Position of gradient changes of feature vector over a sample of testing images from video sequences

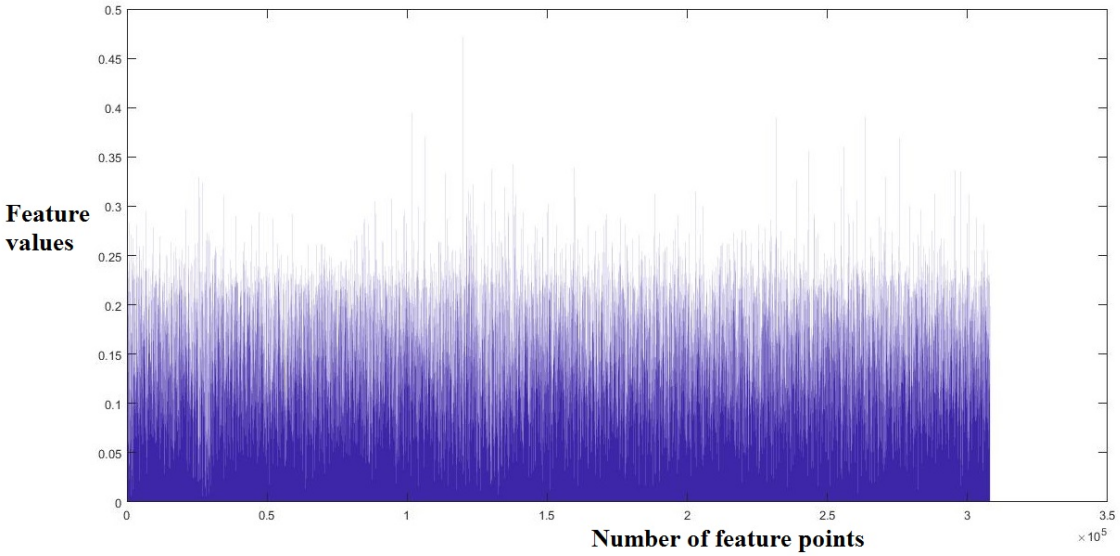


Fig. 21: Feature vector distribution plot of a testing image against no. of features

Figure 22 shows that the accuracy of training and testing for human detection and tracking with polynomial kernel [13]. Here we used hard margin with different values of polynomial kernel. Here we have found the training accuracy is about 1 or 100% and human detection and tracking accuracy is maximum at polynomial value $p = 3$, which is about 92%.

Figure 23 shows that the accuracy of training and testing for human detection and tracking by using soft margin [13], where we use different kernel value of polynomial kernel against C and p value. Here p indicates the different values of polynomial kernel and C be the regularization parameter which is used to control the trade off between the achieving a low training and testing error that is the ability to generate a classifier to unseen data. Here accuracy of human detection and tracking is found to be maximum at polynomial values from $p = 2$ *to* 3 with value of regularization parameter $C = 0$ *to* 2.5 which is about 88%.

Figure 24 shows some samples of human detection and tracking results from different video streams that can be used for human action recognition direction for AVSS design. Here videos are captured in different locations: (i). trunk road on a holiday, (ii). corridoor of a medical cum research centre, (iii). footpath of a trunk road, (iv). wakling zone of an university, (v). a staircase of a buildng (vi). a road in a busy day where strong sunlight as well as shadow of trees are there, (vii). staircase of an another building etc. In this figure each of the rows indicates different video streams and each

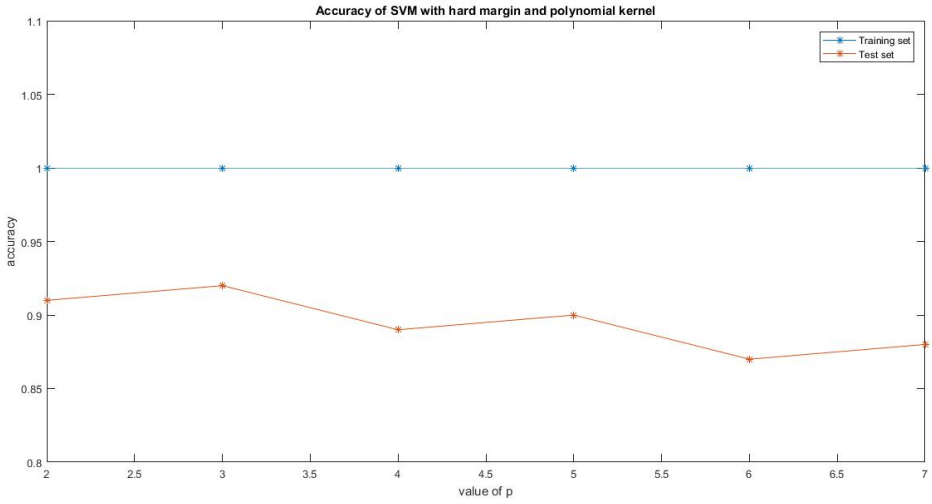


Fig. 22: Accuracy of training and testing for human detection and tracking by using hard margin with different values of polynomial kernel

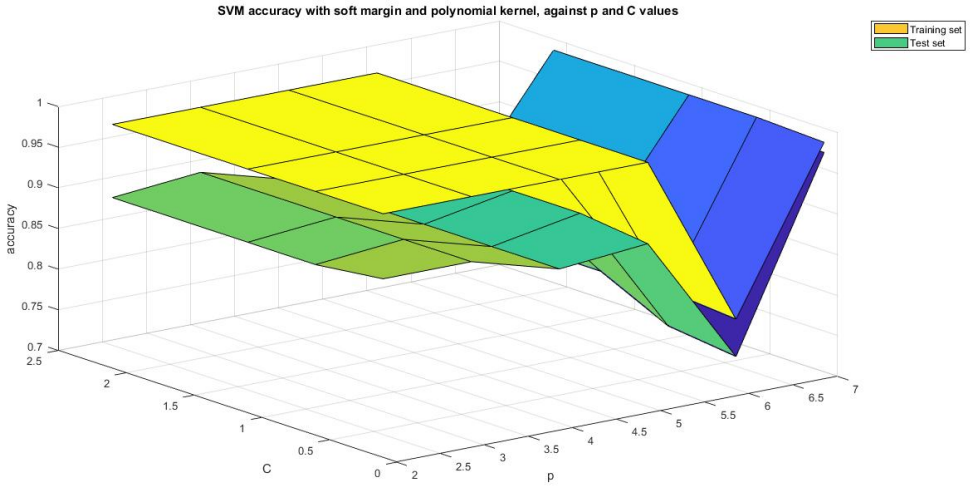


Fig. 23: Accuracy of training and testing for human detection and tracking by using soft margin with different kernel value of polynomial kernel against C and p value

of the columns indicates different frames of that video streams where human detection and tracking takes place for different frames of that video streams: (a). in the first row indicates human detection and tracking placed in different framne numbers of that video, where one human detection and tracking takes place in frame no 20, 368 and 1502 and two human in frame no 530 of video 1, (b). in the second row shows that one human detection and tracking takes place in frame no 2, 260, 502 and 930 of video 2, (c). in the third row shows that two human detection and tracking takes place in frame no 10, 520, 1012 and 1590 of video 5, (d). in the fourth row shows that three human detection and tracking takes place in frame no 10, two human in frame no. 1200 and two paertially occluded as well as more then half way occluded human in frame no. 4800 and 6019 of video 4, (e). in the fifth row one human detection and tracking takes place in frame no 1, 170, 317 and 500 of video 6, (f). in the sixth row one human detection and tracking takes place in frame no 10, two in frame no. 1170, three in frame no. 817 and 1500 of video 9, (g). in the seventh row one human detection and tracking takes place in frame no 5, 360, and two in frame no. 870 and 906 of video 8.

Here in some video streams it is difficult to detect as well as track all of the actual human present. This is mainly due to the (a). the cameraman is moving with the walking people (camera is not fixed positioned) therefore uncertainity takes place, (b). camera can capture video that contains human are out of range (far from the camera positioned), (c). due to the huge sudden variations of the sunlight as well as shadow, (d). due to the fully occluded situations, (e). due to the partialy appearing of human in front of camera etc.

Fig. 24: Some samples of human detection and tracking for action recognition in different videos

## V. Performance evaluation of the proposed system

Performance evaluation/measurement is necessary for detection of object as well as classify different objects. We use confusion matrix as shown in figure 25 for the calculation of those parameters that we needs. Confusion matrix has a link between the actual values and predicted values of object detection.

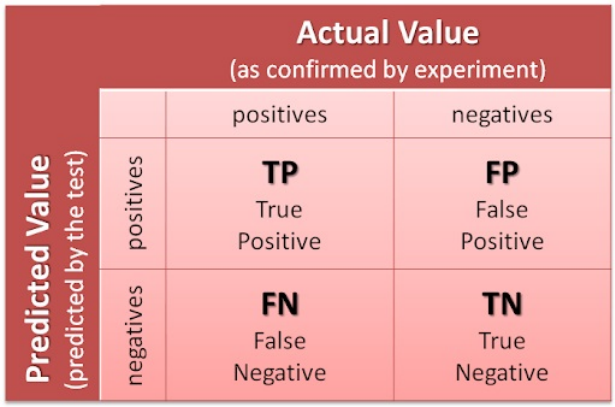


Fig. 25: Schematic of confusion matrix [44]

### A. Precision

Here precision describe the accuracy in picking out a particular type of target (e.g. human or non-human) from data that contains both the human and non-human data. If TP be the true positive rate and FP be the false positive rate of detection, than we have the precision value from the following equation as:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

### B. Recall

Recall describe the success in finding an item in a database (e.g. human or non-human database). Therefore we have the recall value by using following equation:

$$Recall = t_{pr} = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{11}$$

### C. F-measure

To indicate overall performance of a classification system we use the F-measure parameter. We have the equation for F-measure parameter equation as:

$$F - measure = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{12}$$

Table 3: Table for number of frames, positive samples and negative samples against each of the collected videos

| Serial no. | Frames | Positive | Negative |
|---|---|---|---|
| Video 1 | 1567 | 2616 | 3265 |
| Video 2 | 1049 | 1049 | 4430 |
| Video 3 | 496 | 992 | 6944 |
| Video 4 | 6082 | 14160 | 92271 |
| Video 5 | 1656 | 3529 | 19217 |
| Video 6 | 621 | 621 | 2215 |
| Video 7 | 5260 | 15018 | 20968 |
| Video 8 | 943 | 2829 | 9208 |
| Video 9 | 1543 | 14301 | 21960 |
| Video 10 | 3727 | 20124 | 79715 |

We have captured videos for our system in different scenarios. Table 3 shows the total no. of frames of each videos and detected positive and negative samples are present in all of those tested video streams that are captured in different locations. Table 4 shows the true positive, true negative, false positive and false negative values of human detection for all of those tested video streams that are captured in different locations. Table 5 shows the comparison of true positive rate (TPR), true negative rate (TNR), false positive rate (FPR) and false negative rates (FNR) for all of the tested videos. Table 6 shows the values of precision, recalls, F-measures and automatic multiple human detection accuracy for visual surveillance system which shows the individual human detection accuracy from all of those captured image sequences or videos. Table 7 shows the comparison table of performance evaluation for our system with different existing methods, which shows that our system provides better multiple human detection as well as detection and tracking accuracy for visual surveillance system.

Table 4: Table for true positive (TP), true negative (TN), false positive (FP), false negative (FN) values of detection for all of the tested videos

| Sl. no. | TP | TN | FP | FN |
|---|---|---|---|---|
| Video 1 | 2519 | 3150 | 115 | 97 |
| Video 2 | 1024 | 4296 | 134 | 25 |
| Video 3 | 952 | 6890 | 54 | 40 |
| Video 4 | 13657 | 90806 | 1465 | 503 |
| Video 5 | 3399 | 18547 | 670 | 130 |
| Video 6 | 599 | 2181 | 34 | 22 |
| Video 7 | 14758 | 20378 | 590 | 260 |
| Video 8 | 2749 | 9007 | 201 | 80 |
| Video 9 | 13823 | 81499 | 461 | 478 |
| Video 10 | 19438 | 78802 | 913 | 686 |

Table 5: Comparison table for TPR (True Positive Rate), TNR (True Negative Rate), FPR (False Positive Rate) and FNR (False Negative Rate) for all tested videos in %

| Serial no. | TPR | TNR | FPR | FNR |
|---|---|---|---|---|
| Video 1 | 96.30 | 96.50 | 3.50 | 3.70 |
| Video 2 | 97.60 | 97.00 | 3.00 | 2.40 |
| Video 3 | 96.00 | 99.23 | 0.77 | 4.00 |
| Video 4 | 96.40 | 98.40 | 1.60 | 3.60 |
| Video 5 | 96.30 | 96.50 | 3.50 | 3.70 |
| Video 6 | 96.46 | 98.50 | 1.50 | 3.54 |
| Video 7 | 98.30 | 97.20 | 2.80 | 1.70 |
| Video 8 | 97.17 | 97.72 | 2.18 | 2.89 |
| Video 9 | 96.66 | 97.90 | 2.10 | 3.34 |
| Video 10 | 96.60 | 98.85 | 1.15 | 3.40 |

Table 6: Comparison table of precision, recall, f-measure and accuracy for all of the tested videos; all are in percentage %.

| Serial no. | Precision | Recall | F- measure | Accuracy |
|---|---|---|---|---|
| Video 1 | 95.63 | 96.30 | 95.96 | 96.40 |
| Video 2 | 88.40 | 97.62 | 92.78 | 97.10 |
| Video 3 | 96.00 | 94.60 | 95.29 | 98.81 |
| Video 4 | 90.30 | 96.40 | 93.30 | 98.20 |
| Video 5 | 80.00 | 96.30 | 87.40 | 96.48 |
| Video 6 | 97.72 | 96.46 | 97.10 | 98.02 |
| Video 7 | 96.20 | 98.30 | 97.24 | 97.60 |
| Video 8 | 93.20 | 97.17 | 95.15 | 97.67 |
| Video 9 | 96.80 | 96.66 | 96.73 | 97.40 |
| Video 10 | 95.51 | 96.60 | 96.05 | 98.40 |

Table 7: Comparison table for performance evaluation of our method with some existing methods. Here precision, recall, f-measure, human detection rate, human detection and tracking rate are in %. (Here P=Precision, R=Recall, F=F-Measure, DR=Detection Rate, HDTR=Human Detection and Tracking Rate, POSD=Purpose of System Design)

| **Authors** | P | R | F | HDR | HDTR | POSD |
|---|---|---|---|---|---|---|
| Lakshmi et al. [9] | 92 | 96 | — | — | — | Human detection from video |
| Ramya et. al. [10] | 89.11 | 92.83 | 89.25 | — | — | Moving object detection |
| Dalal et. al. [12] | — | — | — | 89 | — | human detection from image |
| Davis et. al. [14] | 100 | 92.73 | — | 95.29 | — | Human detection for robotic tasks |
| Kachoune et. al. [16] | — | — | — | 86 | — | Fast human detection from video |
| Kushwaha et. al. [20] | — | — | — | — | 87.44 | Human detection & tracking |
| Beaugendre et. al. [25] | — | — | — | — | 80 | Human tracking only |
| Ahuja et. al. [29] | — | — | — | 92.20 | — | Pedestrian detection |
| Mishra et. al. [38] | — | — | — | 77.78 | — | Human motion detection |
| Wo et. al. [39] | — | — | — | — | 80 | Multiple human tracking |
| Proposed | 92.98 | 96.64 | 94.70 | 97.61 | 92 | Automatic multiple human detection & tracking for human action recognition |

## VI. Conclusion

Classification of object from video streams are became very challenging in case of machine vision applications. We have mainly focused our human detection and tracking system design to the direction of activity or action recognition for AVSS design. Here we mainly focuse to solve the problem of detection and tracking of human in different lightening conditions, less data requirements for faster system operation as well as handle the occlusion as far we can. Therefore we use median filte as noise removal, graph cut for image segmentation, HOG with different specifications as a feature descriptor, SVM with polynomial kernal as a classifier, and finally particle filter for tracking those of detected human. With this we have found that a 97.61% human detection and about 92% human detection and tracking accuracy. Due to the use of median filter noise part is remove to meet a proper system design. Here we use graph cut segmentation, therefore we have found only less amount of feature vector for the positive training images. So less system memory required as compared to other methods. Also we have segmented the human images so the occlusion can be handled about 88%. The system can work properly in different indoor as well as outdoor environments which can be further used for the design of an AVSS. Further we will extend our work from people detection to human action recognition, through which we can recognize the action is usual or not. This system can be applicable to the design of a proper AVSS design where the system can automatically recognize human, than classify different actions and also check the action is usual or not. If not then the system will generate an alarm through which various crimes and terrorist activities will be catch up. Till now we are using machine learning based technique. In our next work we will focus our work to the field of deep learning platforms for a proper AVSS design.

## Acknowledgment

This work is supported by MHRD-TEQIP- III, Govt. of India.

## References

[1] P. Patidar, M. Gupta, S. Srivastava and A. K. Nagawat "Image de-noising by various filters for different noise," Int. J. Comp. Applications, vol. 9, no. 4, pp. 45-50, Nov. 2010.

[2] X. Benavent, E. Dura, F. Vegara, and J. Domingo, "Mathematical morphology for color images: an image-dependent approach," Mathematical Problems in Engineering, vol. 2012, pp. 1-18, Dec. 2012.

[3] M. Ahmad, "Human motion detection and segmentation from moving image sequences," in Proc. 2008 Int. Conf. Electrical and Computer Engineering, Dhaka, Bangladesh, 2008, pp. 407-411.

[4] T. Ubukata et al., "Fast Human Detection Combining Range Image Segmentation and Local Feature Based Detection," in Proc. 2014 $22^{nd}$ Int. Conf. Pattern Recognition, Stockholm, Sweden, 2014, pp. 4281-4286.

[5] D. V. Thombre, J. H. Nirmal and L. Das, "Human detection and tracking using image segmentation and kalman filter," in Proc. 2009 Int. Conf. Intelligent Agent & Multi-Agent Systems, Chennai, India, 2009, pp. 1-5.

[6] F. Hafiz, O. Khalifa, A. A. Shafie and M. H. Ali, "Foreground segmentation-based human detection with shadow removal," in Proc. Int. Conf. Computer and Communication Engineering (ICCCE'10), Kuala Lumpur, 2010, pp. 1-6.

[7] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in Proc. 8th IEEE Int. Conf. on Computer Vision (ICCV 2001), Vancouver, BC, Canada, 2001, pp. 105-112.

[8] Q. Zhang and K. N. Ngan, "Segmentation and Tracking Multiple Objects Under Occlusion From Multiview Video," in IEEE Trans. on Image Processing, vol. 20, no. 11, pp. 3308-3313, Nov. 2011.

[9] N. D. Lakshmi,Y. M. Latha and A. Damodaram, "Silhouette extraction of a human body based on fusion of HOG and graph-cut segmentation in dynamic background," in Proc. $3^{rd}$ Int. Conf. Computational Intelligence and Information Technology (CIIT 2013), Mumbai, India, 2013, pp. 527-531.

[10] A. Ramya and P. Raviraj, "Performance evaluation of detecting moving objects using graph cut segmentation," in Proc. 2014 Int. Conf. Green Computing Communication and Electrical Engineering (ICGCCEE), Coimbatore, India, 2014, pp. 1-6.

[11] S. R. Kharabe, P. S. Hanwate, D. S. Gaikwad and K. P. kaliyamurthie, "Human image segmentation," in Proc. 2017 Int. Conf. Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1-3.

[12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for human detection," in Proc. 2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893.

[13] C. P. Diehl and G. Cauwenberghs, "SVM incremental learning, adaptation and optimization," in Proc. of the Int. Joint Conf. Neural Networks, Portland, OR, 2003, pp. 2685-2690.

[14] M. Davis and F. Sahin, "HOG feature human detection system," in Proc. 2016 IEEE Int. Conf. Systems, Man, and Cybernetics (SMC), Budapest, 2016, pp. 2878-2883.

[15] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng, "Fast human detection using a cascade of Histograms of Oriented Gradients," in Proc. 2006 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006, pp. 1491-1498.

[16] M. Kachouane, S. Sahki, M. Lakrouf and N. Ouadah, "HOG based fast human detection," in Proc. 2012 $24^{th}$ Int. Conf. Microelectronics (ICM), Algiers, 2012, pp. 1-4.

[17] W. J. Park, D. H. Kim, Suryanto, C. G. Lyuh, T. M. Roh and S. J. Ko, "Fast human detection using selective block-based HOG-LBP," in Proc. 2012 19th IEEE Int. Conf. Image Processing, Orlando, FL, 2012, pp. 601-604.

[18] A. E. Bell, "Robust feature vector for efficient human detection," in Proc. 2013 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, 2013, pp. 1-5.

[19] T. Zhao and R. Nevatia, "Tracking multiple human in complex situations," IEEE Trans. on pattern analysis and machine intelligence, vol. 26, no. 9, pp. 1208-1221, Sep. 2004.

[20] A. K. S. Kushwaha, C. M. Sharma, M. Khare, R. K. Srivastava and A. Khare, "Automatic multiple human detection and tracking for visual surveillance system," in Proc. 2012 Int. Conf. on Informatics, Electronics & Vision (ICIEV), Dhaka, 2012, pp. 326-331.

[21] A. Adam, E. Rivlin, I. Shimshoni and D. Reinitz, "Robust real time unusual event detection using multiple fixed location monitor," IEEE Trans. on pattern analysis and machine intelligence,vol. 30, no. 3, pp. 555-560, March 2008.

[22] A. Galata, N. Johnson, and D. Hogg, "Learning behaviour models of human activity," in Proc. British Machine Vision Conf., 1999, pp. 12-22.

[23] A. K. R. Chowdhury and R. Chellappa, "A factorization approach for activity recognition," in Proc. 2003 Conf. on Computer Vision and Pattern Recognition Workshop, Madison, Wisconsin, USA, 2003, pp. 41-41.

[24] O. Chomat and J. L. Crowley, "Probabilistic recognition of activity using local appearance," in Proc. 1999 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 1999, pp. 104-109.

[25] A. Beaugendre, H. Miyano, E. Ishidera and S. Goto, "Human tracking system for automatic video surveillance with particle filter," in Proc. 2010 IEEE Asia Pacific Conf. on Circuits and Systems, Kuala Lumpur, 2010, pp. 152-155.

[26] J. Saboune and R. Laganiere, "People detection and tracking using the explorative particle filtering," in Proc. 2009 IEEE 12th Int. Conf. Computer Vision Workshops, ICCV Workshops, Kyoto, 2009, pp. 1298-1305.

[27] X. Li, K. Wang, W. Wang and Y. Li, " A multiple object tracking method using kalman filter," in Proc. 2010 IEEE Int. Conf. Information and Automation, Harbin, 2010, pp. 1862-1866.

[28] S. Rahimi, A. Aghagolzadeh and H. Seyedarabi, "Human detection and tracking using new features combination in particle filter framework," 2013 8th Iranian Conf. Machine Vision and Image Processing (MVIP), Zanjan, 2013, pp. 349-354.

[29] S. Ahuja and P. Pandey (2015, June). Pedestrian detection using HOG features. IIT Delhi, New Delhi, India. [Online]. Available: http://sarthakahuja.org/public/docs/report_ped_

detection.pdf.

[30] W. Lao, J. Han and P. H. n. De With, "Automatic video-based human motion analyzer for consumer surveillance system," in IEEE Trans. Consumer Electronics, vol. 55, no. 2, pp. 591-598, May 2009.

[31] K. Seemanthini, S.S. Manjunath, "Human Detection and Tracking using HOG for Action Recognition," Procedia Computer Science, Vol. 132, 2018, pp. 1317-1326.

[32] F. Angelini, Z. Fu, S. A. Velastin, J. A. Chambers and S. M. Naqvi, "3D-Hog Embedding Frameworks for Single and Multi-Viewpoints Action Recognition Based on Human Silhouettes," 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 4219-4223.

[33] C. Thurau, "Behavior histograms for action recognition and human detection," In Proc. $2^{nd}$ conf. Human motion: understanding, modeling, capture and animation. Springer-Verlag, Berlin, Heidelberg, 2007, pp. 299–312.

[34] S. P. Sahoo, R. Silambarasi and S. Ari, "Fusion of histogram based features for Human Action Recognition," in Proc. 2019 $5^{th}$ Int. Conf. on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 1012-1016.

[35] Jagadeesh. B and C. M. Patil, "Video Based Human Activity Detection, Recognition and Classification of actions using SVM", TMLAI, vol. 6, no. 6, p. 22, Jan. 2019.

[36] A. B. Sargano, A. Plamen, and Z. Habib. 2016. "Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines," Applied Sciences, vol. 6, no. 10, Oct. 2016, pp. 309-313.

[37] N. Johnson and D. Hogg,"Learning the distribution of object trajectories for event recognition," in Proc. of the $6^{th}$ British conf. Machine vision, 1995, pp. 583-592.

[38] S. K. Mishra and K. S. Bhagat, "Human motion detection and video surveillance using MATLAB," Int. J. Scientific Engineering and Research (IJER), vol. 3, no. 7, pp. 154-157, July 2015.

[39] B. Wu and R. Nevatia, "Tracking of multiple humans in meetings," in Proc. 2006 Conf. Computer Vision and Pattern Recognition Workshop (CVPRW'06), New York, NY, USA, 2006, pp. 143-143.

[40] Y. Chen, G. Liang, K. K. Lee and Y. Xu, "Abnormal behaviour detection by multi-SVM based bayesian network," in Proc. 2007 Int. Conf. Information Acquisition, Seogwipo-si, 2007, pp. 298-303.

[41] D. T. Nguyen and K. R. Park,"Enhanced gender recognition system using an improved histogram of oriented gradient (HOG) feature from quality assessment of visible light and thermal images of the human body," Sensors, vol. 16, no. 7, July, 2016.

[42] G. Levi, "A short introduction to descriptors," Aug. 18, 2013. [Online]. Available: https://gilscvblog.com/2013/08/18/a-short-introduction-to-descriptors. [Accessed Jan. 29, 2021].

[43] L. Konwar, A. K. Talukdar and K. K. Sarma, " Design of Automatic Visual Surveillance System - Methods and Approaches," BULLETIN OF SCIENTIFIC RESEARCH (Previously J. of Assam Science Society), vol. 60, no. 1, ISSN 0587-1921 pp. 41-60, 2019-2020.

[44] Confusion matrix, Available: https://alearningaday.blog/2016/09/14/confusion-matrix/. [Accessed Jan. 29, 2021].

[45] P. Xiao, M. Yuan, X. Zhang, X. Feng and Y. Guo, "Cosegmentation for Object-Based Building Change Detection From High-Resolution Remotely Sensed Images," in IEEE Trans. Geoscience and Remote Sensing, vol. 55, no. 3, pp. 1587-1603, March 2017.