ARTIFICIAL INTELLIGENCE FOR HUMAN BEHAVIOR ANALYSIS

Divyashree M H[1], C.S. Shivaraj[2]

[1]Electrical and Electronics, The National Institute of Engineering, Mysuru, Karnataka, India
[2]Electrical and Electronics, The National Institute of Engineering, Mysuru, Karnataka, India

***Abstract** - As Natural intelligence (NI) is exhibited by humans and other animals, Artificial intelligence (AI) is programmed into machines. AI research field is described as the study of intelligent devices like a Robot, computer etc that recognize its surrounding environment and takes appropriate actions. In the field of video surveillance system, moving object is a prominent area of research under computer vision. This is not effortless work as continual distortion of entity taking place during motion. Any entity in motion has various accredit in spatial and temporal spaces. In temporal space entity varies in motion rate where as in spatial space entity differs in size .The main focus of this work is detection and identification of people. The video datasets of group of people are considered in order to identify humans and track humans in crowd scene. Background subtraction technique is utilized to detect humans. To extract features Histogram of Oriented Gradient feature descriptor technique is applied. Support Vector Machine (SVM) classifier method is made use to recognize human activity performed.*

***Keywords**:-* **Natural intelligence (NI), Artificial intelligence (AI), Support Vector Machine (SVM), Background subtraction (BS), Threshold (T).**

I. INTRODUCTION

Artificial Intelligence in Computer vision is used to learn different ways to analyze, re-build and to comprehend 3-dimensional pictures from its two-dimensional scenes determined on the actual link of structures in attendance in the particular video. It predominantly comprises of techniques to obtain, understand, examine and process the digital images. Video processing is prominent in social gatherings, country border, banks, Sports Stadium, Offices, Airports and shopping complexes. The issue of human being detection, tracking and activity recognition has gained importance in the field of computer vision. The identification and tracking the moving objects and activity recognition of these bodies in video surveillance system is a major task. In recent times this is used for various artificial intelligent video managing and monitoring systems. Applications includes finding out abnormal activities, giving security using surveillance video, patient control unit, sports video, traffic management etc.

This is a well organized and effective approach used mainly to find motion based event and activity identification in the video set recorded. In some conditions more amounts of actions may sometimes show variations due to less video quality, altering background, overlapping situation, different human view-points, and disturbance in background and also because of many changing entities in the surrounding. Recognition of Human being activity is mainly utilized for human interactions with each other as it gives knowledge about identity of the people, their way behaving, impression of their personality etc. It is widely use in interaction between computer robotics and human beings which uses depiction of many people's behavior. All these need a system which identifies different kind of activity.

II. RELATED WORK

Human detection and their activity recognition is a very active research area in Artificial Intelligence through computer vision especially monitoring suspicious activity. Fangbing Zhang, Lisong Wei,and others[1] proposed the concept called as Cube surface modeling for human detection in crowd. In this paper, they have proposed a novel real-time and reliable human detection system. The human detection problem is solved by presenting a novel cube surface model captured by a binocular stereo vision camera. They first proposed a cube surface model to estimate the 3D background cubes in the surveillance area and then developed a shadow-free strategy for cube surface model updating. Thereafter, they present a shadow weighted clustering method to efficiently search for human as well as remove false images. But this method failed to recognize the Human activity.

Tao Ji, Leibo Liu [2]and others came up with a concept called Fast and Efficient Integration of Human Upper-Body Detection and Orientation Estimation in RGB-D Video .This paper presents a novel integrated approach for human upper-body detection and orientation estimation. This approach does not restrict the feature type so that any combination of RGB and depth features can work seamlessly in the proposed framework. Human upper-body detection and orientation estimations are incorporated using a random forest model. But this paper failed to detect the human action performed.

Kuei-Chung Chang and Po-Kai Liu[3] proposed a tracking algorithm for head detection in crowd scenes. This paper proposed an approach to detect multiple heads, which can be applied in smart human tracking system. The computing resource of this kind of applications is so high that it is not applicable in embedded platforms. So, this paper proposes a parallel design to enhance the performance of the approach such that it can be more applicable in embedded platforms. This algorithm deals only with head detection technique and also fails to track long trajectory and the actions performed by the people.

Yanan Zhang, Hongyu Wang and Fang Xu [4] proposed Object detection and recognition is the premise and foundation for intelligent service robot to understand the surrounding environment and make intelligent decisions. In this paper, aiming at the accuracy and real-time performance of object detection and recognition of service robot in complex scenes, an end to end object detection and recognition algorithm based on deep learning is proposed. The algorithm has both good accuracy but failed to recognize the activities performed by humans.

Hiromasa Taada and others [5] proposed Human tracking in crowded scenes using target information at previous frames. In this paper the method compare the target region with similar regions at current frame. In addition, they also compare the target region at current and previous frames. The probability of uncommon colors at current and previous frame is reduced thereby improving the tracing activity. This only holds good in tracing the humans in video but fails to recognize their activity performed by them.

MyoThida and others [6] proposed two different modeling techniques for human activity detection and tracking in crowded scenes. In order to know about particular motion patterns in crowded scenes first technique called macroscopic modeling technique was used and the second type of technique is microscopic modeling, mainly depends on analyzing the video path of moving entities. However, this modeling technique captures normal movement of crowd scenes but fails to detect details of individual person actions.

III. OVERVIEW PROPOSED SYSTEM

The proposed work includes detection, tracking and activity recognition of many people from a video datasets.

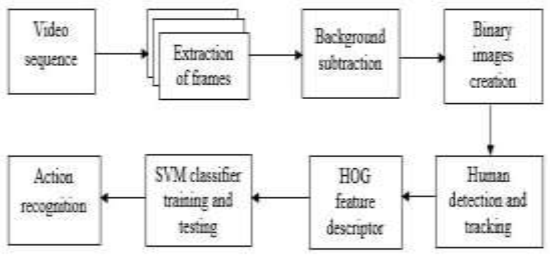

Figure 3.1 Architecture for HAR

3.1 Video dataset

The datasets in recorded video are taken where people performs various actions in different backgrounds with various video formats.

3.2 Pre-processing

The chief motive of pre-processing stage is to device the data for feature extraction and to remove the noise.

3.3 Frames Extraction

It is one of the prominent stages where the video inputted is converted into many frames. The value of total number of frames is based on the recorded video length. Further the converted frames can be utilized for detection, processing, and activity identification.

3.4 Background Subtraction

The foreground is obtained for processing using this method. This is a prominent step in order to determine the people in motion in the video taken. In people identification the region of interest is human in motion in its foreground. Hence with the help of this technique moving people can be extracted from its background image. It identifies human in moment by subtracting the reference image (background image considered under static background condition) and current image frame.

A vigorous Background subtraction method must manage changes in lighting and also the considerable long term changes in video scene. This inspection utilizes the function (x, y, t), here (x, y) denotes location of the pixel in x and y coordinates and time dimension (t) in the video sequence. A easy way to implement this method is to consider a background image as a reference frame (denoted by 'B') and a frames acquired at t (time interval) represented as 'C(t)'.By utilizing simple mathematical calculations it is feasible to determine the person just by using image subtraction method for every picture element present in C(t), numbers of current images of the picture element is shown by P [C(t)] and reduced with its respective value of pixel in certain place of the background scene is denoted as P [B].

The difference image obtained will show some of the intensity components for the picture element place which are diverted in following two frame positions considered for background subtraction. This approach shows good outcome when all the foreground picture objects are in motion and all the background objects are constant. In order to make subtraction outcome better a threshold (T) is utilized on this difference photo.

The mathematical form for thresholding is written as:

$$P\,[C(t)]P\,[B]\,]> T$$

3.5 Human Activity Recognition

This is one of the significant technologies as it can be exercised in real-time scenarios. So far investigation is concentrated on simple activity identification carried out by people like running, walking, hand waving, etc. The main intention of designing HAR system is to analyze automatically existing events, motion and to get the needed context from the data captured. The Figure 3.2 denotes the basic framework for HAR system.

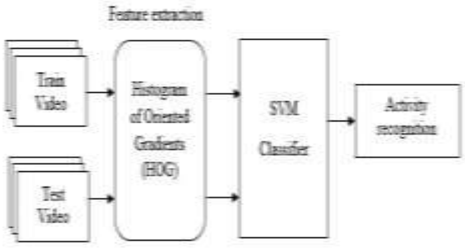

Figure 3.2 Structure of HAR

3.6 HOG feature descriptor

This approach makes use of diminishing the number of resources using various approaches which outline a huge set of data. The technique Feature extraction is utilized to determine, segregate several desired features in the digitized image. Once after the creation of binary images, next task is to extract the features. Features are the interesting parts of an image which are stored in a compact vector form called feature vectors. Proposed work uses HOG approach is made use to get required information from account the tedious aspect of gradient within the confined position of the considered image and is segregated to tiny attach portion called as cells and for pixels present inside every cell a HOG direction are organized. Subsequently the descriptor focuses on the acquired histograms. The attributes extracted called HOG feature vector are put to use for instructing the SVM.

Generally the attributes got will be in '1' by the 'M vector', where 'M' denotes HOG feature range. Consequently, the obtained data could be extensively utilized in categories, identify and track. Feature extraction of area of people in motion a using HOG approach is as indicated in figure 3.4. In order to get HOG attributes the input dataset video is taken into account, where moving human being is first recognized and only region in motion will be taken out. Once after this, HOG technique is made use to get the attributes present of moving people and for every features obtained histogram will the plotted. To obtain the attributes, videos are taken as the input which is given to gradient computation block in which direction and magnitude will be computed, after which these resulted values are sent to 'Gradient Vote' block proceeded by normalization is done and at the last phase all HOGs will be computed.

3.6.1. Gradient computation

For each of the pixel (x, y) the magnitude m(x, y) and direction (x, y) is found. Assuming that pixel which is to be computed is situated at coordinates (x, y) and f(x, y) is its luminance value. The Gradients in x-axis and y-axis is denoted by 'f_x(x, y)' and 'f_y(x, y)'. The gradient values are determines as follows

$$f_x(x; y) = f(x + 1; y)f(x \quad 1; y)$$

$$f_y(x; y) = f(x; y + 1)f(x; y \quad 1)$$

Magnitude m(x; y) and the direction (x; y) is computed as:

$$m(x,\ y) = \quad f_x(x; y)^2 + f_y(x; y)^2$$

$$(x, y) = \tan \ \frac{f_x(x; y)}{f_y(x; y)}$$

3.6.2 Gradient vote

After magnitude 'm(x, y)' and direction '(x, y)' are obtained, for all pixels Gradient vote is computed which in turn gives orientation histogram. The orientation is fairly spaced with '0', to '180'. The weight of every pixel is given by:

$$/= (n + 0{:}5)b \ \ (x; y)$$

Where 'n' indicates a bin where direction (x; y) and the value of 'b' is used denote the total number of bins. Two neighboring bins values are incremented to reduce aliasing. Thus m_a and m_b is given as:

$$m_a = (1) \quad m\,(x; y)$$

$$m_b = m\,(x; y)$$

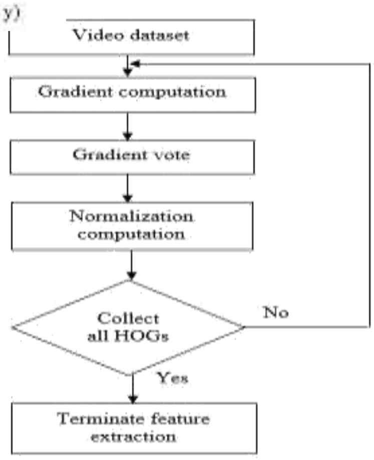

Figure 3.3 Dataflow for HOG Feature Extraction

3.6.3 Normalization computation

At the Last stage this technique is applied by combining all the histograms which is a part of particular block.

3.7 SVM Classifier

SVM is an organized learning model. The learning algorithm is used to inspect the data which is need for classification.

IV. IMPLEMENTATION

Artificial Intelligence for Human Behavior Analysis is implemented using 'MATLAB'. The initial step in people identification and HAR (Human activity recognition) is taking the video datasets. The datasets consists humans doing moments like stand, walk, punch, Handshake, hug, kick and fallback are taken. Implementation of the 'HAR' begins with extraction of frame. The frame extraction, which starts with acquiring information about size of the video file, memory, quality, resolution, etc. by giving command called 'aviinfo'. Utilizing the command 'videoreader' video is read and then number of frames in video is generated using 'numframes'. After obtaining all the total frames loop execution starts until it reach end of frames, after that in video file frames are read.

Once after obtaining the frames next operation is to convert these frames to image files by making use of command 'frame2im'.In the end the converted image will be saved. The rate of Frame extraction is '30' frames per second. This step is necessary as videos cannot be directly processed .Figure 4.1 denotes the flowchart of HAR system. Later, the BS technique is utilized to identify the humans in motion. In this method a background image will be considered in which each frame will be subtracted by this background image in order to get foreground images which contains the humans region. Foreground 'RGB' image is converted into 'Gray' scale images. In order to remove the noise components in the resultant '2-D median filtering' is utilized.

Once after done with removal of noise 'Gray scale' images will be converted to 'Binary images' of zeros and ones, where binary 0 is used which represents absence of humans and binary 1 represents white region with human presence. Hence, to extract any moving people and objects present in the video binary image creation is very useful. After this process the dilation operation is carried out on obtained binary images.

Measurements mainly includes area which gives actual number of pixel present in the image region, bounding box which represents a small rectangle box in which individual human region exist, centroid which determines the center pixel of each detected human and many more. On conducting all the tasks individual and also group of people will be detected. In order to recognize human action 'SVM classifier' is made used.

In order to create the classifier major step is to select training sets. In present work about seven different kinds of training folders are created. It contains the action activities like walk, stand, handshake, kick, punch, hug and fall back. The SVM classifier needs N(N-1)/2 training classifier where N denotes the number of training folders. 'One to one strategy' is made use for classification. While testing, when a sample is given as input to the classifier if the output is positive, that group belongs to class A else belongs to class B.

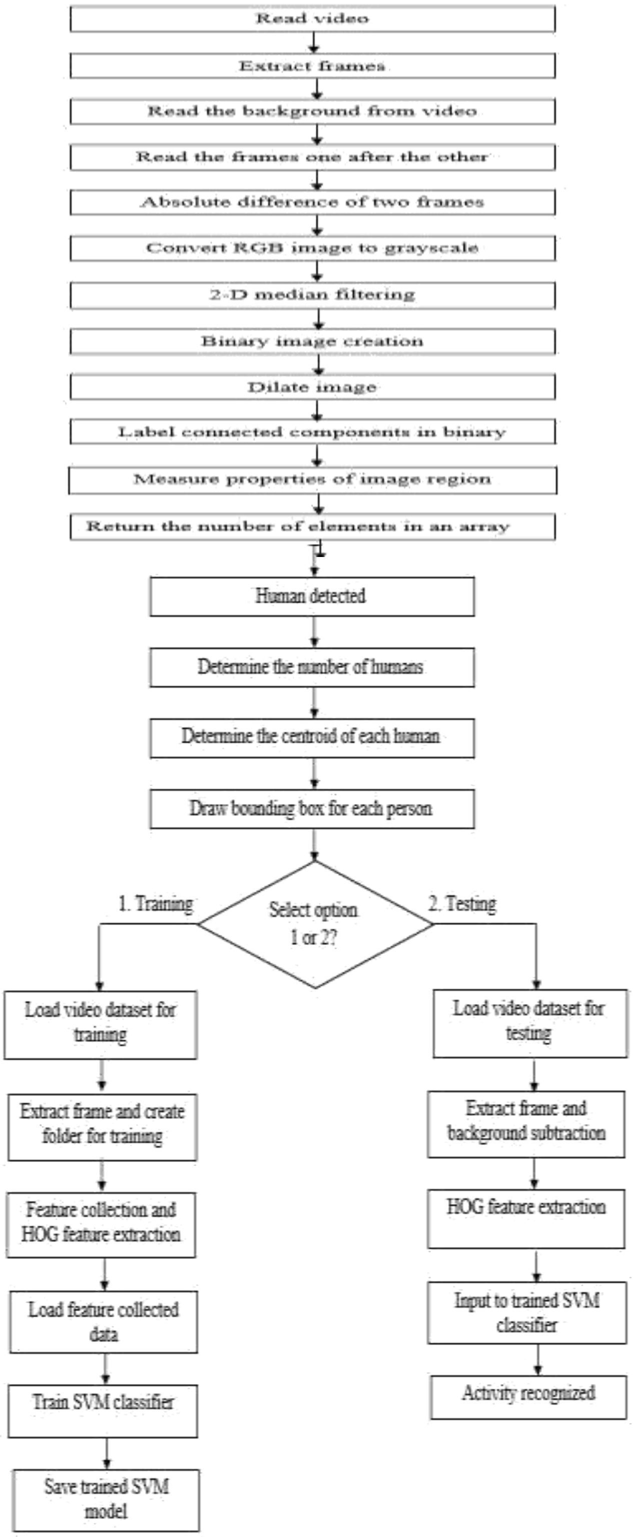
Read video
Extract frames
Read the background from video
Read the frames one after the other
Absolute difference of two frames
Convert RGB image to grayscale
2-D median filtering
Binary image creation
Dilate image
Label connected components in binary
Measure properties of image region
Return the number of elements in an array
Human detected
Determine the number of humans
Determine the centroid of each human
Draw bounding box for each person
1. Training
Select option 1 or 2?
2. Testing
Load video dataset for training
Extract frame and create folder for training
Feature collection and HOG feature extraction
Load feature collected data
Train SVM classifier
Save trained SVM model
Load video dataset for testing
Extract frame and background subtraction
HOG feature extraction
Input to trained SVM classifier
Activity recognized

Figure 4.1 Flowchart of HAR

V. RESULTS

The experiment conducted by video datasets which are taken under various kinds of backgrounds, occlusions and expressions. This is implemented using 'MATLAB' software.

5.1 Test video 1

This video is UT-Interaction type dataset which is of High quality. In this video there are 4 people doing 5 different actions like handshake, punch, stand, walk, and kick. The duration of the recorded video is of 13 seconds and having memory size of 1.21MB. The Pixel resolution of video is 720 480 having 29 frames per second. The following figures show various activities recognized by the system.

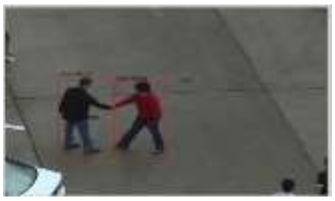

Figure 5.1: HAR of 2 people with handshake activities

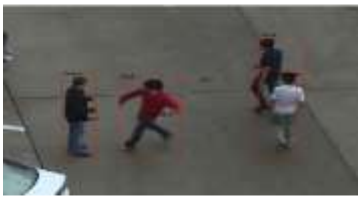

Figure 5.2: HAR of 4 people with stand, kick and walk actions

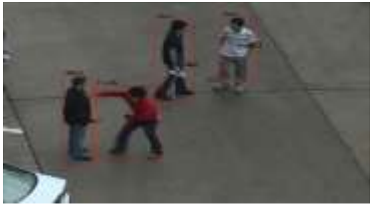

Figure 5.3: HAR of 4 people with stand, punch and walk actions.

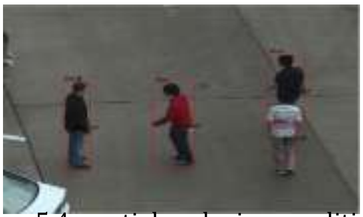

Figure 5.4: partial occlusion condition

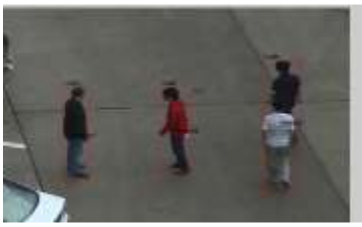

Figure 5.5: complete occlusion condition

Table 5.1: Generalized Tabulation for first test video

Frame number	Action	Person
1	Handshake	2
2	Stand	1
2	Kick	1
2	Walk	2
3	Stand	1
3	Punch	1
3	Walk	2
4	Stand	1
4	Walk	3
5	Stand	1
5	Walk	2

5.2 Test video 2:-

This video dataset consist of 5 people performing seven varieties of activities like walk, stand, hug, handshake, punch, kick and fallback action)The video length is 27- seconds with size- 2.97 MB and 720 *480 pixel resolution.

The Frame rate of the video is 25fps (frames per second). Figures show the experimental results of the test video 2.

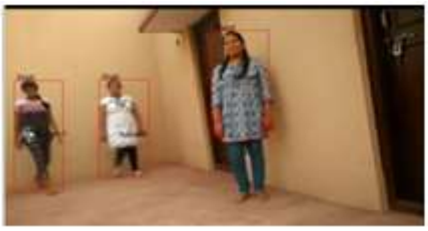

Figure 5.6: HAR of 3 persons with walk and stand activities

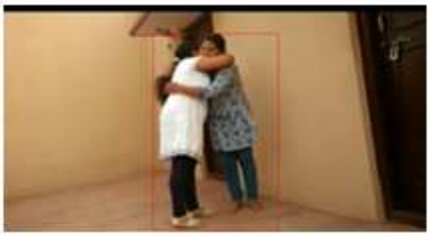

Figure 5.7: HAR of 2 persons with hug activity

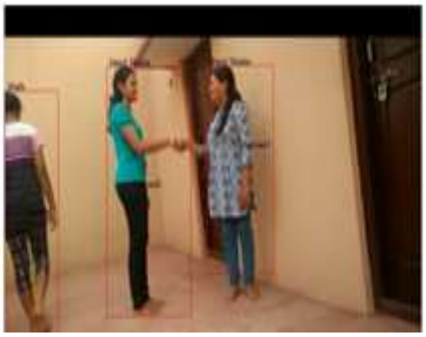

Figure 5.8: HAR of 3 persons with walk and hand shake activities

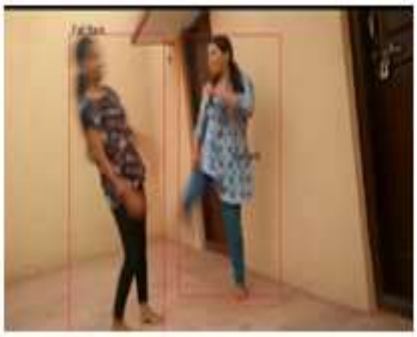

Figure 5.9: HAR of 2 persons with kick and fall back activities

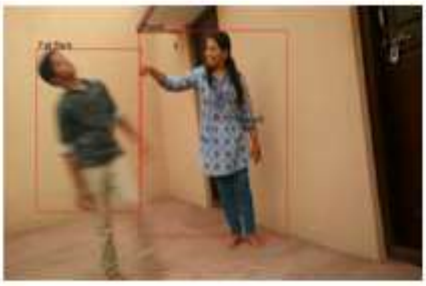

Figure 5.10: HAR of 2 persons with punch and fall back activities

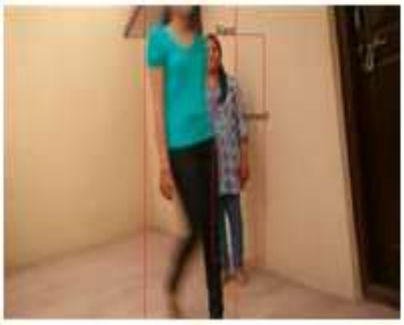

Figure 5.11: shows the partial occlusion condition

Table 5.2: Generalized tabulation for test video 2

Frame Number	Action	Person
1	Stand	1
1	Walk	2
2	Handshake	1
3	Walk	1
3	Handshake	2
4	Kick	1
4	Fall back	1
5	Punch	1
5	Fall back	1
6	Walk	1
6	Stand	1

5.3 Test video 3:-

This video-dataset comprises of 4 humans performing 4 varieties of activities like walk, stand, handshake and punch. The Video duration is 24 seconds having size 2.64MB and [720X480] pixel resolution. Frame rate of video is of 25 frames/ second. Figures show the outcome of test-video3.

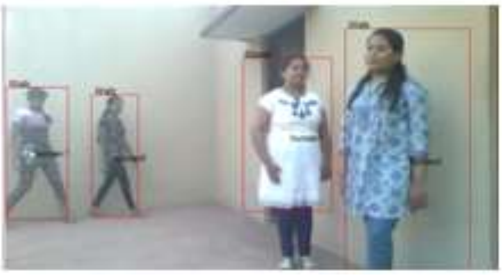

Figure 5.12: HAR of 4 persons with walk and stand activities

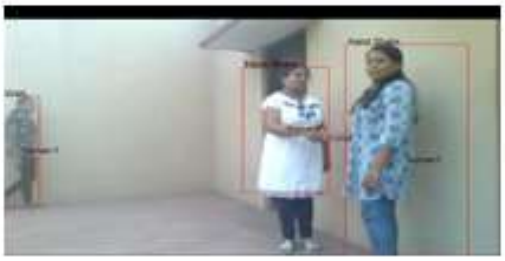

Figure 5.13: HAR of 3 persons with handshake and walk activities

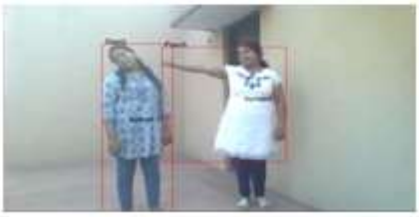

Figure 5.14: HAR of 2 persons with punch activity

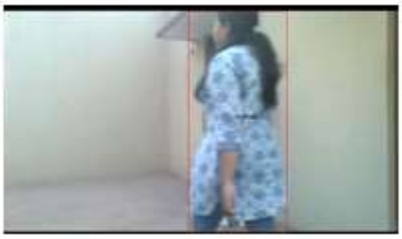

Figure 5.15: Occlusion condition with complete overlap

Table 5.3: Generalized tabulation for test video 3

Frame Number	Action	Person
1	Stand	1
1	Walk	3
2	Walk	1
2	Handshake	2
3	Stand	1
3	Punch	1
4	Walk	1

VI. APPLICATIONS AND FUTURE SCOPE

Artificial Intelligence for human detection and behavior recognition finds application in various areas. Some of them are listed below:-

- Video surveillance to identify the illegal activities.
- To notify and provide security during theft or robbery (for example in banks, malls, hospitals and also in country border).
- To secure public from illegal terrorist activities. (In Mumbai blast at famous 'Taj hotel', the terrorist attacked and captured the place for about four days. They carried out shooting innocent people and made the entire world to fear. If this technique is implemented there, the shoot activity can be considered illegal and the system can automatically inform to the security persons. Within a couple of minutes the situation can easily be controlled).
- Gaming and sport applications to recognize the activities being performed. (It can be used in foot-ball game to make the robots give commentary by considering the video.)

In the field of Artificial Intelligence through computer vision People detection and recognition in crowd is an important and challenging and task. The Future work aims to recognize many kinds of actions like eating, recognition of face, talking, mood identification considering the facial expression and other factors which can be done by understanding the body gesture of each individual person in the video. The Future Artificial Intelligence to detect humans and to analyze their behavior can be imported to a Robot.

REFERENCES

[1] Jing Li, Fangbing Zhang, Lisong Wei1, Tao Yang and Zhongzhen Li1, Cube surface modeling for human detection in crowd ",International Conference on Multimedia and Expo (ICME) 2017

[2] Tao Ji,Leibo Liu ,Wenping Zhu, Jinghe Wei and Shaojun Wei, Fast and Efficient Integration of Human Upper-Body Detection and Orientation Estimation in RGB-D Video,"2017 IEEE 9th International Conference .

[3] Kuei-Chung Chang and Po-Kai Liu, Design and Optimization of Multiple Head Detection for Embedded System, 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE 2017)

[4] Yanan Zhang, Hongyu Wang and Fang Xu Object detection and of intelligent service robot based on deep learning, IEEE 8th International Conference on CIS & RAM, Ningbo, China.

[5] Hiromasa Takada ,Kazuhiro hotta and PranamJanney ,Human tracking in crowded scenes using target information at previous frames,2016 23rd International CONFERENCE ON Pttern Recognition(ICPR),Cancun center, Mexico, Dec 4-8-2016

[6] M. Thida, Y. L. Yong, P. Climent-Perez, A literature review on video analytics of crowded scenes in Intelligent Multimedia Surveillance Springer, 2013, pp. 17–36.

BIOGRAPHIES

Divyashree M H received BE degree in Electrical and Electronics from Vidhyavikas Institute of Engineering, Mysuru. She is currently an Mtech student in Computer applications in industrial drives in The National Institute of Engineering, Mysuru, India.

C.S. Shivaraj is an Assistant Professor in the department of Electrical and Electronics Engineering in National Institute of Engineering, Mysuru. He has received his BE , Mtech Degree from VTU, Belgaum.