CHAPTER 1

INTRODUCTION

1.1 Background

Due to an increase in crime and terrorists activity, the concern for public security increases. Video Surveillance Systems are now of great use for monitoring and management of public areas. Video Surveillance System emphasize on Human Detection and its behaviour. In recent years, Computer Vision has gained importance in the field of Human Detection. The main goal and challenge in computer vision is to detect the object in an image and track that object. Recent Researches in the field of Computer Vision has increased its focus on observing humans, understanding their appearance and activities that are kept under continuous surveillance, intelligent control and human computer interaction**.** Humans can be differentiated from other objects by its shape, ratio of human width and height. Human has almost a constant height to width ratio of their body and so they can be easily distinguished from other objects. Considering the case of human detection, the challenge is to detect people with number of poses with variable clothing and appearance in complex backgrounds as well as occluded objects. The problem is further compounded by ambiguous background clutter, changing of lighting conditions, shadowing and self-shadowing. The understanding of Human Activities has various applications like surveillance, animations for gaming and movies, activity recognition and bio-mechanical analysis of actions for sports. The primary step for monitoring people activities is by classifying the movement of various objects. We can summarize a Human Detection and Tracking System as follows. The first and the foremost requirement is an appearance model which is needed to be constructed correctly. Then, detection is performed based on this model to get the correct position of the person. Finally, tracking is done by continuous detection of person in consecutive video frames to obtain the correct trajectory of person. Researches on Human Detection is not only based on detection of single person but also based on detection of multiple persons. Many detectors are proposed to detect humans from still images. Characteristic information of human is extracted (such as Gradient Information which has a powerful discriminative property and can represent the shape of the human body well) to distinguish human from other object. If human

detection from video is required, motion information can be used. Motion caused by human is different from motion caused by other objects. Robustness is an important factor for detection. Many systems cannot read motions accurately or otherwise cannot function optimally due to factors such as background or lighting changes. In addition, they do not always properly recognize motions made against noisy backgrounds and cluttered scenes. Depending on different applications the requirements for detection and tracking may differ significantly, e.g. detection of single object is much easier as compared to detection of multiple objects.

Tracking can be defined as the estimation of trajectory of moving objects in sequence of images. Automatic detection and tracking of moving object is very important task for human computer interface, video communication/expression and security and surveillance system application and so on. Various imaging techniques for detection, tracking and identification of the moving objects have been proposed by many researchers. in the real world scenarios, that is a very challenging task due to the interference of noise, clutters, occlusions, illumination variations and dynamic changes of the object and the background appearance in the complex scene; a quite variety of tracking methods have been proposed to tackle these difficulties in decades (*Yilmaz et al.* [1]), which can be roughly divided into two categories: the deterministic method and the statistical methods.

Depending on different applications the requirements for detection and tracking may differ significantly, e.g. detection of single object is much easier as compared to detection of multiple objects. Accuracy is a significant issue in Multiple Human Tracking. The trajectory of humans is important for surveillance. A multiple people tracking algorithm should make sure that the system tracks the same person under different situations such as temporary people occlusion. There is always a trade-off between precision and computing time because if we want to increase the precision the computing time increases and if we want to decrease the computing time the precision also decreases. In almost all the circumstances, robustness and accuracy needs to be satisfactory then algorithm optimization and speed of the system are taken into consideration for improvement of performance of the system.

Human detection in Computer Vision indicates a category of methods that estimate locations of human bodies in images or video frames. Objects to be located can be full

or part of (usually upper half) human bodies. Methods of human detection can be divided into two categories: Signal-processing-based methods, which classify objects by matching specific spatial information, and machine-learning-based methods, which statistically learn models and classify objects. Human tracking methods are applied to estimate movement of human bodies in frame sequences.Movement of human bodies can be achieved by analysing differences between frames, or finding out connections between located human bodies in consecutive frames.

The main challenges of Human Tracking are:-

- **Inter-object occlusion: -** This occurs when a part of an object is hidden behind another. It is directly proportional to the crowd density.
- **Self-occlusion: -** Suppose a person talks in a mobile phone then the hand of the person occludes itself. This type of occlusion is short term.
- **Size of The Visible Region: -** The Size of the Visible Region of the object is inversely proportional to the density of the crowd. So, detecting and tracking object is difficult in case of a dense crowd.
- **Appearance ambiguity: -** Appearance tends to be less distinguished when the objects appear small.

The tracking methods are categorized into following categories:-

Region Based Tracking(Wren et al. **[2]*; McKenna*[3]*):-*** Region-based tracking algorithms track objects according to variations of the image regions corresponding to the moving objects. For these algorithms, the background image is maintained dynamically and motion regions are usually detected by subtracting the background from the current image. Wrenused small blob features to track a single human in an indoor environment. The pixels which belong to the human body are assigned to the different body part's blobs. The moving human is successfully tracked by tracking each small blob.

Active Contour Based Tracking **(*Paragois&Deriche*** **[4]*)*:-** The basic idea is to allow a contour to deform such that the energy function is minimized to produce the desired segmentation. They are categorized into 2 types:-

Edge Based Contour Models: - To identify Object Boundaries they utilize image gradients. One major advantage of this method is that no constraint is placed on the

image. So, the foreground and background can be heterogeneous and the correct segmentation can be achieved. The major drawback of this method is that it is very sensitive to image noise and depends on initial curve placement.

Region Based Contour Models: **-** The foreground and the background are statistically modelled and the optimum energy is found where the model best fits the image. In this model, the various image regions is of constant intensity is assumed.

The advantages of region-based approaches over the edge based methods include robustness against initial curve placement and insensitivity to noise. The techniques that attempt to model regions using statistical features are not ideal for heterogeneous object segmentation.

Feature Based Tracking(Schiele, **[5]*; Coifman et al.,*** **[6]*):-*** Feature Based Object Tracking consists of Feature Extraction and Feature Correspondence. A Feature Point in one image may have similar points in another image in another image which results in ambiguity in Feature Correspondence. To remove this ambiguity algorithms perform exhaustive search or compute the correlation of pixels over large windows. So, in this case the computational complexity increase. In the method of *Tomsai and Kanade*[7], they used small windows to track the translational motion of objects by minimizing the error and thus the complexity is reduced. But, for longer sequence their approach loses a significant percentage of tracked points.

Model Based Tracking(Koller**[8]*):-*** The main challenges of Video Tracking are the Real Time Videos are often recorded in unfavorable environments for example like in low lights or variable weather like in the rains and so on. These factors often cause undesired luminance and contrast variations in videos produced by optical cameras (e.g. the object entering dark or shadowy areas).

The trajectory of humans is important for surveillance. A multiple people tracking algorithm should make sure that the system tracks the same person under different situations such as temporary people occlusion. There is always a trade-off between precision and computing time because if we want to increase the precision the computing time increases and if we want to decrease the computing time the precision also decreases. In almost all the circumstances, robustness and accuracy needs to be satisfactory then algorithm optimization and speed of the system are taken into consideration for improvement of performance of the system.

1.2 Objectives

This thesis aims to develop a robust pedestrian detection framework which is able to detect pedestrians in challenging situations where the background color matches with the skin color of the human. The light intensity levels also varies in the scene to add to the ambiguity. To achieve this goal a number of sub-objective has been set.

- To identify the most important feature to describe human which is invariant to illumination.
- To use the de-noising filers to remove the unwanted noise in the scene.
- To track if a human is in the scene whether it is static or moving in each frames of a video.

1.3 Motivations and Presuppositions of the Thesis

The motivations of the thesis are to advance the area of human detection, to take a background which matches with the skin color of the human and to investigate a method that improve performance of human detection.

The presuppositions which are assumed in our work throughout the thesis are as follows:

- Images and videos are captured by conventional 2D RGB cameras.
- Both image and video data used is uncompressed. Compressed data need to be decompressed to be applied to the system. Hence compressed images are converted to raw data and compressed videos are converted to frame sequences.
- We have considered the noise issue at present but if the background becomes different then further planning for removal of noise is required.
- Video cameras are fixed, which means that frame differences caused by movement of cameras can be ignored.
- We focus on processing videos containing one entire human body. Although multiple bodies can be processed in our system.
- The algorithm is tested on various situations. It is tested on a self-made database which has a background colour similar to that of Human skin colour.

This thesis makes the following main contributions:-

- Uses the Aspect Ratio of Human as a Feature to detect and track Humans.

- Removes if any noise is present in the background.
- Detects a moving Human Body as well as a Stationary Human Body.
- Calculate the Detection percentage of Human in the whole video sequence.

1.4 Overview of the System Proposed in the Thesis

This thesis presents a 2D human motion analysis system which is used to detect human bodies.If the input is a video, we design to process and estimate the position of human in each frame separately, remove the unwanted objects or noise in the background, detect if human is present in the scene and post-process the estimated results to achieve a final detection percentage.

The flow diagram is shown below:-

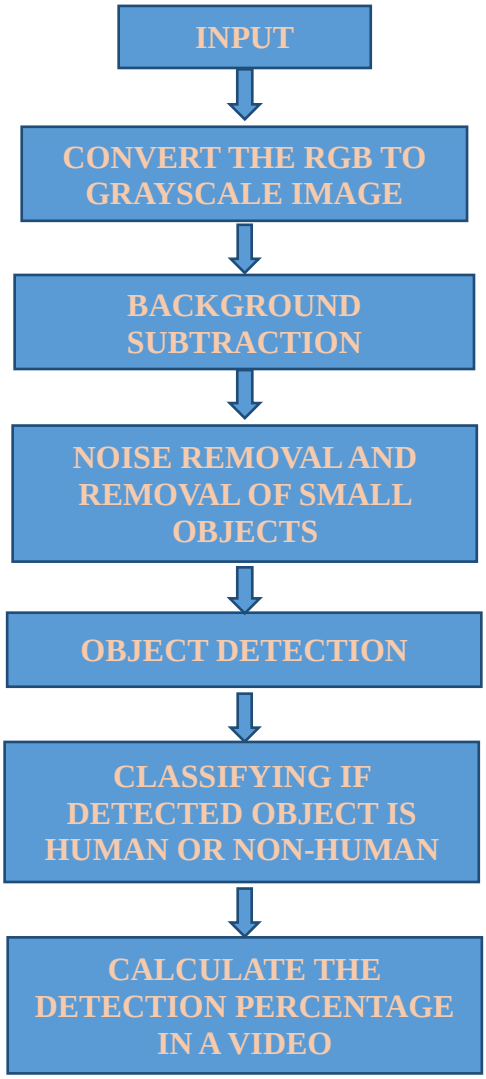


Fig. 1 System Flowchart

1.5 Organization of the Thesis

After the Introductory Chapter, Rest of the thesis is organized as follows:-

- **Chapter 2** conducts a literature review of Human Detection Research. In this chapter, current approaches and techniques are surveyed. Various parameters of Human detection is also surveyed in this Chapter
- **Chapter 3** provides an overview of the Proposed Human Detection Framework along with the whole process flow of the Algorithm. Each of the steps are presented in detail.
- **Chapter 4** provides with the Experimental Results of the Algorithm with various Datasets. It also gives a comparative analysis of our method with other existing methods and also shows the Detection Ratio of the Algorithm.
- **Chapter 5** summarizes the Thesis with Conclusion, Advantages, Limitations and Future Work.

CHAPTER 2

LITERATURE REVIEW

Wren et al paper shows the detection of humans by the method of Background Subtraction. In this technique, the foreground is first found and then this foreground object is classified into humans, birds, etc. The Background is modelled by a Gaussian distribution in YUV Space at each pixel. This background is continuously updated. The background is fixed while the foreground is moving so the Spatial Components continuously changes. These Spatial Components are constantly estimated in each frame by a Kalman Filter [9].

Dalal and Triggs[10]in their paper found a very efficient approach for detecting humans. A number of human poses are added as a training data and selected a detection window of size normally 64*128 pixels. Then in order to normalize the image, they used gamma normalization where the square root of the intensity or its log value is taken. The gradients of the image were computed and the magnitude and orientation of these gradients were found. The detection window is splitted into dense grid and in each grid we have cells. In each cell, the pixel gradient its orientation and magnitude are figured out. Histogram of orientation in these cells was created and the histogram is weighted by magnitude of the gradient. So, Histogram of Orientation of the Gradients was created. Several cells in the neighbourhood have been combined to create a block which normalizes the image that reduces the effect of contrast variation. Histogram of Oriented Gradients was collected over the detection window and put them one after another in a feature vector and then they used a learning algorithm like Linear Support Vector Machine is used in order to complete the process of detection. The Detection efficiency is very good for this method but this method is too slow to be used for real time applications.

Zhu [11]speeded up the system of Dalal and Trrigs by combining Cascade of Rejecters Approach with the Histogram of Oriented Gradient Feature. HoGs of variable block sizes are used to capture Human Feature. They used the Adaboost Algorithm to identify appropriate set of blocks from a large set of blocks. For a 320*280 image, the

system possesses a speed of 5 to 30 frames per second. This approach is relatively faster than the technique used by Dalal and Trrigs. For a Sparse scan, i.e. 800 windows per image, the time required by Dalal and Triggs's method is 500 msec while for the approach by *Zhu* is 25 msec. Similarly for a Dense Scan i.e. the time for Dalal and Triggs's experiment is 7 sec while that for Zhu's approach it is 100 msec. The training time is much more and also the memory required in Zhu's approach is more. Real Time Detection was achieved by combining HoG features and Adaboost algorithm. This method improved the speed significantly. The detection performance was also quite similar to the previous methods but False Positives per Window (FPPW) increased.

Hui-Xing Jia[12]combined Viola's Object Detection Method and Histogram of Oriented Gradient features to detect humans efficiently. In this method, the Haar Features on Viola's Framework is substituted by Histogram of Oriented Gradient Feature. So, this experiment shows that both the speed of Viola's Object Detection Method as well as the discriminative power of HoG is preserved. The results show less False Positives per Window and less False Positive Rate at a comparable speed. The system is much slower than Viola's system which used a Haar Adaboost system in sparse scan. But the speed of this system is comparable in case of Dense Scan while it is much faster than Dalal's method of HOG-SVM Bootstrapping system in case of sparse scan. However, the system is much slower than Dalal's method in case of Dense Scan because the feature evaluation is much simplified in this method.

Li Zhihui[13] in his research has shown a method of detecting humans with a good detecting speed as well as reducing the False Alarm rate. First a scanning window is used to detect human by the combination of HoG features and Adaboost Algorithm. If the result of Human Detection is positive, the colour histogram of images is calculated as a feature near the head. If this detection is positive then again an Adaboost classifier is used to filter the positive window. The calculation of Colour Histogram reduces the False Positives per Window. The Adaboost Algorithm is used to improve the speed of the system. This algorithm can be said to be a compromise between better detection results and lower detection time.

Xu Fen, Li Jie[14] implemented the Adaboost Algorithm on a DSP Platform for Human Detection. Haar like features as weak classifiers were used and a strong classifier is obtained from it. A number of positive and negative samples are trained and

a Cascade of Boosted Classifiers is obtained. The experiment is performed with Adaboost Human Detection Algorithm on DM642. DM642 is a high performance media processor dedicated to video applications. In this research work, 18 stages of Cascade Adaboost human detection algorithm is tested using the INRIA [15] person image library. TheDetection Ratio by this method is up to 93% but the ratio of false detection is also very high. To reduce False Detection Rate, the number of poses in the training stage must be increased. The main disadvantage of this method is the execution time of this DSP based Adaboost Algorithm for human detection is significant and can't be used for real time applications.

Changyan Li[16]recently found an efficient method of human detection and tracking algorithm where they achieved an overall detection rate above 80%. This experiment reduced the detection time and improved human detection and tracking accuracy. This system detected humans based on Improved HoG i.e. *Zhu's* technique (Cascade of Rejecters Technique). Here Improved HoG is applied to save detection time or to apply the system for real time applications. Tracking also provides some information about detection so making the detection time faster. Then, Kalman Filter is used to predict and estimate the detected human in the next frame. So, the accuracy of human tracking is improved. At first, people in the video are detected by Zhu's method of improved HoG. Some useful information, like location, etc for Kalman Filter initialization has been noted. Kalman Filter is employed for predicting people's position. The Kalman Filter can also get updated with the speed of moving human constantly, so enhancing the accuracy. When detection is done on the next frame, the scale of detection is narrowed or we can rather tell that the human is being tracked. The distance between the tracker and the human is being reduced frame by frame. So, the human is detected as well as tracked by this method. This method reduces the detection time greatly because the Kalman Filter updates the position of the human in each frame. It also improves the detection efficiency. The experiment is performed by using both simple HoG as well as Improved HoG. The detection time by using normal HoG is 1600 ms in each frame while by Improved HoG i.e. the method by *Zhu* the detection time is 50msec for each frame. The detection speed and tracking results are quite good. The only disadvantage of the method is it cannot track multiple people in the video.

Heewook Jung[17] detected humans on bicycle. Bicycles are more difficult to be detected because its appearance changes with movement. HoG feature and Real

Adaboost Algorithm is used in this method. Occlusions are a great cause for the decrement of detection efficiency. The next position of the human on the bicycle is decided by Particle Filtering. The movement of bicycles is non-linear sometimes. The Particle Filter is used because it is robust to non-linear systems. This method can avoid accidents because it can track objects i.e. in this case bicycles which is being occluded also. This method can also detect the bicycle's driving direction because three classifiers learn respectively the training samples of the driving direction of bicycle (front, left and right). The HoG models of Bicycles are trained via a Real Adaboost Algorithm to detect the Bicycles. When HoG as a feature is used to detect the bicycles then many windows containing a single bicycle is obtained. These windows have been merged by Mean Shift Clustering and Nearest Neighbor Algorithm. With Real Outdoor Scenes, this experiment detected humans on a bicycle with a precision of 90% and with the False Positive Rate as low as 10%. The driving direction is detected with a precision of 94% and a False Positive of 6%. The tracking results are also very good with a precision of 88% and a False Positive Rate of 12%.

Li Li, XU Jining[18] presented a new method for moving human detection based on static camera. Moving Human Detection Methods can generally be classified into 2 categories- first is the Temporal Difference which doesn't give a clear picture of the contours of moving object and second is the Optical Flow whose computation time is quite high and also the noise rejection quality is quite poor. In this experiment, multi-dimensional Gaussian distribution model is used for each pixel background. Background Subtraction is then implemented to detect the motion of humans. The shadow of detected human is eliminated in HSV space. This would reduce the False Positives per Window by a great deal. Then background noises and disturbances are eliminated by morphological operations. In this experiment, the learning rate α is kept fixed at 0.02. This works well when the background is fixed or rather short time applications but if we consider long time applications or changed background, then α has to be modified. The computing time of this method is more because objects are detected in the first step and the shadows are detected and eliminated in the next step.

Xiaohui Liu [19] presented a Robust Approach for Multiple Human Detection, Tracking and Identification. They used Zhu's method of Improved HoG for Human Detection. Color Histogram and SIFT Feature [20] are used to increase the robustness for inter-frame matching. In case of Multiple Human Detection, a person's location

could be very close, overlapping or even might be occluded. The tracking might become invalid. A temporal differencing algorithm is employed to find the moving region. The searching becomes easier by specifying the moving regions. The searching time is also vastly reduced. For Tracking, Kalman Filtering is used and for Human Identification, both Colour Histograms and SIFT Features are used. SIFT feature generates and updates local feature of humans while the Colour Histogram is a statistical feature for the appearance of humans. So, SIFT feature is integrated with Colour Histogram Feature to produce a signature for different individuals. INRIA data set is used to train the Improved HoG Classifier and used 1000 positive and 1500 negative images. A Detection rate of 98% and false positive rate of 1% for static images is obtained by this method. Public videos from PET 2009 dataset are used. This experiment worked well for multiple people and the time taken is also very less. Average Frame Rate of this method is 9.7frames per second.

Schwartz [21] provided texture and colour information with edge-based features to obtain a rich descriptor set. This resulted in a high dimensional feature space. Support Vector Machines (SVM) cannot walk on such a high dimensional feature space. The Partial Least Squares (PLS) [22] Analysis was used over this high dimensional descriptor set. The number of samples in the training data set is much smaller as compared to other methods. Partial Least Square Method is used as a class dimensionality reduction tool. PLS is used to project the high dimensionality feature vectors to a subspace of very low dimension where SVM can be used as a classifier. Humans in a standing position have characteristic of strong edges present along the body of human, uniformity in clothing texture as well as the uniformity of the ground texture. Humans were discriminated from background by edges, colour and texture. Low level feature by HoG descriptors were captured with additional colour information, and the texture features were computed from the co-occurrence matrix. The HoG descriptor with colour information along with the co-occurrence matrix were concatenated and analysed to reduce the high dimensional feature space to a lower one. Then a simple and efficient classifier like SVM is used to classify whether the object is a human or a non-human. The results shown by this method outperformed other previous methods of human detection. The main feature of this paper was the integration of edge-based feature, texture measure and colour information which

resulted in a high dimensional feature space and then converting it to a low dimensional feature space by the method of Partial Least Squares Method.

Yah-Li Hou[23] detected multiple human in a complex situation like in a crowded scene. They first extracted the KLT features from the original image. The filtering of foreground mask was done and thus the background feature points are removed. The feature points from human beings remained. The clustering of these feature points are done with some prior knowledge of human size. We can use the human body ratio also (ratio of height and width). The GMM (Gaussian Mixture Model) [24] is used for foreground extraction. The morphological closing operation was performed on this extracted foreground to obtain a foreground mask. The clustering process is then done on this foreground mask by the EM (Expectation Maximization) [25] based Clustering method.

CHAPTER 3

THE METHODOLOGY

3.1 Image Representation and Acquisition

A digital image is a two-dimensional (2D) discrete signal. Such signals are represented as functions of two independent variables - for example, brightness function of two spatial variables. A monochrome digital image $f(x, y)$ is a 2D array of luminance values. Each element of the array is called a pel (picture element), or morecommonly a pixel.A colour digital image is represented by a triplet of values, one for each of the colour channels. The individual colour values are universally 8-bit values, resulting in a total of 3 bytes per pixel. There are number of alternative methods of storing the image data. The most widely used are pixel-interleaved (or meshed) and colour-interleaved (or planar) formats. Row-wise or column-wise interleaving methods are less frequent. In a pixel-interleaved format, every image pixel is represented by a list of three values.

The imaging sensors plays important role in the image acquisition. The structure and operation of the eye is very similar to an electronic camera. Both are based on two major components: a lens assembly, and an imaging sensor. The lens assembly captures a portion of the light flowing out from an object, and focuses it onto the imaging sensor. The imaging sensor transforms the pattern of light into a video signal, either electronic or neural. Focus means there is a one-to one mapping of every point on the object with a corresponding point on the screen. Most of the incident light photons is reflected in random direction, only a small portion of these reflected photons will pass through the lens.

The most common image sensor used in electronic cameras is the Charge Coupled Device (CCD). The CCD is an integrated circuit that replaced most vacuum tube cameras in the

1980s.The heart of the CCD is a thin wafer of silicon, typically about 1cm square. However, there is a new CMOS image sensor that promises to eventually become the image sensor of choice. Both CCD and CMOS image sensors capture light on a grid of small pixels on their surfaces. The output of most sensors is a continuous electrical signal whose amplitude and spatial behaviour are related to the physical phenomenon being sensed. To create a digital image, we need to convert the continuous sensed data into digital form. This is done by Sampling followed by Quantization.To convert the imageinto digital form, we have to sample thefunction in both coordinates and in amplitude. Digitizing the coordinate values is calledSampling. Digitizing the amplitude values is called quantization. The result of sampling and Quantization is a matrix of real numbers. Each element of this matrix array is called an image
Element.

The first digital cameras used CCD (Charged Coupling Devices) to turn images from analog light signals into digital pixels. They are made through a special manufacturing process that allows the conversion to take place in the chip without any sort of distortion. This creates high quality sensors which produces excellent images. They are more expensive than the newer CMOS image sensors.

CMOS (Complementary Metal Oxide Semiconductor) chips use transistors at each pixel to move the charge through traditional wires. This offers flexibility because each pixel is treated individually. Traditional manufacturing processes are used to make CMOS.As they are easier to produce, CMOS sensors are cheaper than CCD sensors. The drop in price of the Digital Cameras is due to the use of CMOS Sensor. CMOS technology came after CCD Sensors and are cheaper to manufacture.

The main difference between the CMOS and CCD Sensor is listed below:-

CCD	**CMOS**
Creates High Quality Images with low Noise.	Creates Images with more Noise.
More Sensitive to Light.	Less Sensitive to Light.
More Power Consumption.	Less Power Consumption.
Higher Manufacturing Cost.	Less Manufacturing Cost.
Vertical Streaking is present in case of Video Mode.	Vertical Streaking is absent

Table 3.1 Difference between CCD and CMOS Cameras

Presently, CCD Sensors produces higher quality images at high resolutions while the CMOS Cameras are catching up with the quality and resolution of CCD sensors. In terms of Power Consumption, CMOS Cameras have a much longer battery life as compared to the CCD Sensors.

3.2 Overview of the Method

The Video usedin the project is a RGB Video. The video is broken into frames and the each frame is converted into a Grayscale Image. Background Subtraction is performed in this Grayscale Image. Since edges reflect strong intensity change therefore we try to find out the edges by converting them to binary image and by some threshold operation and the image is filled as shown below.

Fig. 3.1The Background and Frame No. 29 of Video Dataset A

Fig.3.2The Grayscale Image of Background and Frame No. 29 of Video DatasetA

Fig.3.3The Background Subtracted Image

The next step is to convert this image to a binary image by thresholding and remove the unwanted noise components by some morphological operations.

Fig 3.4The Noisy and the De-Noised Image

So, these are the steps of pre-processing of the data where each frame is processed one by one and background subtraction is performed in each of these frames. The noise and the unwanted small objects is then successfully removed by the morphological operations. However if we change the scene then the value of the filters used is to be changed to a certain extent so that we can get the de-noised image.

Another method is the advanced background subtraction. It only extracts the objects in motion. The i^{th} frame is subtracted with $i\text{-}n^{th}$ frame to get the object in motion. Suppose there is motion in these n frames then these motion can be extracted. This method is called Frame Differencing.

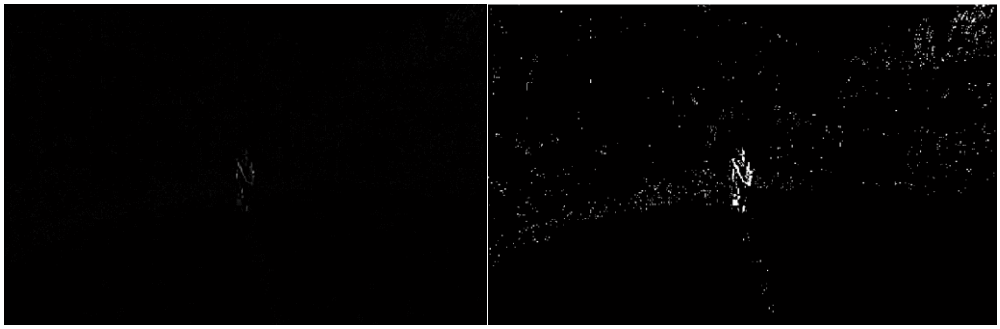


Fig. 3.5 Frame Differenced Image and Binary Image

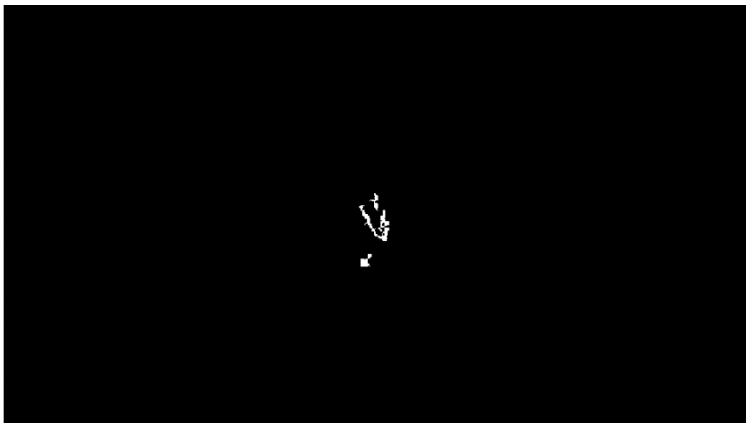


Fig. 3.6 De-noised Image after Morphological Operation

So, Frame differencing results shows better accuracy than the original Background Subtraction Method. The Fig. 7 shows the only movin0g object that is a human while Fig. 5 shows the De-noised Image of the Background Subtraction Method which shows all the objects which are present along with the noise. So, we can conclude that the Frame Differencing Method is more accurate but if any stationary object is to be detected then this method fails i.e. if in a scene an object is stationary from the first frame itself then the method of Frame Differencing method fails, we have to use background subtraction in this case.

Background subtraction can be defined as the technique in the fields of image processing and computer vision where the foreground of an image is extracted for further processing (object recognition etc.). Generally, Regions of Interest in an image are objects (humans, cars, text etc.) in its foreground. After image pre-processing (which may include image de-noising etc.) object localisation is required which may make use of this technique. Background subtraction is widely used for detecting moving objects in videos from static cameras. The motive of this approach is detecting

the moving objects from the difference between the current frame and a reference frame, often called “background image”, or “background model.

In this project, the background subtraction is used in the Grayscale colour space and is further converted into a Binary Image by some thresholding operations.Binary images may contain numerous imperfections. In particular, the binary regions produced by simple thresholding are distorted by noise and texture. Morphological image processing removes these imperfections by accounting for the form and structure of the image.Morphological image processing is a collection of non-linear operations related to the shape or morphology of features in an image. Morphological operations rely only on the relative ordering of pixel values, not on their numerical values, and therefore are especially suited to the processing of binary images. Morphological techniques probe an image with a small shape or template called a structuring element. The structuring element is positioned at all possible locations in the image and it is compared with the corresponding neighbourhood of pixels. Some operations test whether the element "fits" within the neighbourhood, while others test whether it "hits" or intersects the neighbourhood.

After the Morphological operations in each frame we check whether there is any object present or not. If there are objects, then we are calculating the w/l ratio i.e. the Aspect Ratio of the object and if the w/l ratio is coming in the range of 0.3 to 0.75 then we can conclude that the object is a Pedestrian. So we can detect a pedestrian and also a stationary human. If we change the aspect ratio to a value about 2 to 3.5 then we can detect vehicles. So, with a small modification we can detect the vehicles as well. However, in case of occlusion this method doesn’t work well.

3.3 Description of the System

3.3.1 Pre-processing:-3 different videos were taken in a background where the color of the background matches with the skin color of the human and with a variable lighting condition as shown in fig. 3.7.The frames are then processed one by one and is converted into grayscale images. Background Subtraction is performed in these grayscale images. A Robust background subtraction algorithm must handle lighting changes, repetitive motions from clutter and long-term scene changes.

The main aim of the approach is detecting the foreground objects as the difference between the current frame and an image of the scene's static background.

$$\left| frame_i - background_i \right| > Th$$

Where Th is the threshold value.

Fig. 3.7Variable Lighting Condition

After subtracting the background frame from the current frame a threshold Th is applied to the absolute difference to get the foreground mask.

The background frame, current frame and the foreground mask is shown below:-

Fig. 3.8 (a) The Background, (b) The Current Frame and (c) The Foreground Mask

Another method of extraction of the foreground mask is by frame differencing where the difference between the current frame and the previous frame is found. This method is best suitable for estimating whether there is motion or not. But if we find the difference between the $i^{th}$and $i+1^{th}$ frame, there is a very small difference so the moving object cannot be determined. We have to take frames which will give movement say for general case frame no. 15 and 5 rather than consecutive frames.After frame differencing, the grayscale image is converted into binary image by thresholding operations but there will be some unwanted objects that would be present which is removed by different morphological operations.

3.3.2 Image Segmentation: - Image Segmentation is used to separate all the moving objects. This is done in 2 steps. First, the unwanted noise is removed by morphological operations and then Connected Component Analysis is done to find the number of moving objects present in the frame and separate them. The steps are described below:-

(a) ***Unwanted Noise Removal:*** **-**The field of mathematical morphology contributes a wide range of operators to image processing. The operators are useful for the analysis of binary images and common usages include edge detection, noise removal, image enhancement and image segmentation. Morphological techniques typically probe an image with a small shape or template known as a structuring element. The structuring element is positioned at all possible locations in the image and it is compared with the

corresponding neighborhood of pixels. Morphological operations differ in how they carry out this comparison.

An image can be defined as a set or collection of either continuous or discrete co-ordinates. The set corresponds to the pixels that belongs to the objects in the image.

The object ***A*** consists of those pixels that share some common property:-

$$\mathbf{A = \{\alpha \mid property\ (\alpha) == TRUE\}}$$

The background of A is given by A^c (the complement of A) which is defined as those elements that are not in A:

$$\mathbf{A^C = \{\alpha \mid \alpha \notin A\}}$$

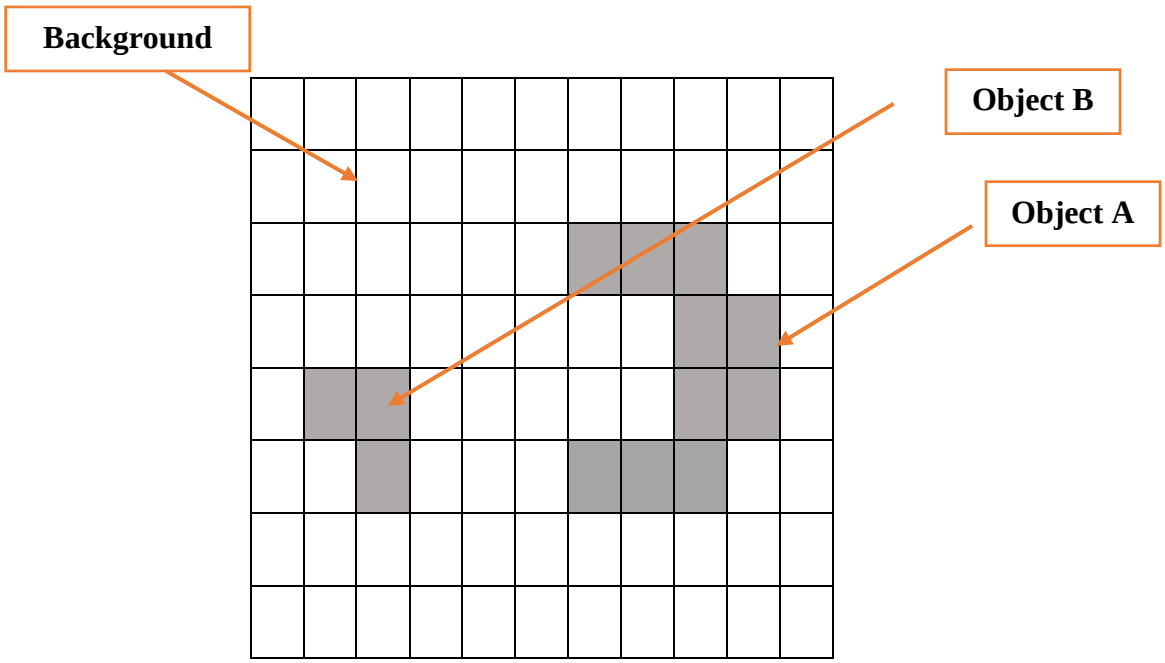


Fig. 3.9A binary image containing two object sets A and B

Some of the fundamental morphological operations are used in the project, like Opening of an Image. To understand the working of opening of the image, we first have to know the definition of structuring element. The structuring element consists of a pattern specified as the coordinates of a number of discrete points relative to some origin. Normally Cartesian coordinates are used and so a convenient way of representing the element is as a small image on a rectangular grid.The **structuring element** is a set of point coordinates (although it is often represented as a binary image). It differs from the

input image coordinate set in that it is normally much smaller, and its coordinate origin is often not in a corner, so that some coordinate elements will have negative values. The figure below is an example of a structuring element where the dark colour is the origin

1	1	1
1	1	1
1	1	1

Fig. 3.10 An example of a structuring element

In the simplest structuring elements used with binary images for operations such as erosion, the elements only have one value, conveniently represented as a one. More complicated elements, such as those used with **thinning** or **grayscale morphological operations**, may have other pixel values.

The Fundamental Morphological Operations are:-

- ***Erosion:* -** The basic operation is to erode away the boundaries of regions of foreground pixels. So, the areas of foreground pixels shrink in size, and holes within those areas become larger.The erosion operator consists of two inputs.The first is the image which is to be eroded. The second is the structuring element.The structuring element determines the precise effect of the erosion on the input image.We consider each of theforeground pixels in the input imageto compute the erosion of a binary input image by a structuring element. . For each foreground pixel (which we will call the *input pixel*) we superimpose the structuring element on top of the input image so that the origin of the structuring element coincides with the input pixel coordinates.If for every pixel in the structuring element, the corresponding pixel in the image underneath is a foreground pixel, then the input pixel is left as it is but if any of the corresponding pixels in the image are background then the input pixel is set to background value. Erosion operation is useful in removing noise and also holes in foreground or background. The expression for Erosion is:-

$$A - B = \{z | (B)_z \subseteq A\}$$

Fig. 3.11 Erosion Effect

- ***Dilation:*** -The basic effect of Dilation on a binary image is to enlarge the boundaries of regions of foreground pixels. So, the areas of foreground pixels grow in size while holes within those regions become smaller.The dilation operator consists of two inputs.The first is the image which is to be dilated. The second is the structuring element.The structuring element determines the precise effect of the dilation on the input image.We consider each of thebackground pixels in the input image to compute the dilation of a binary input image by a structuring element. For each background pixel (*input pixel*), the structuring element is superimposed element on top of the input image such that the input pixel positioncoincides with the origin of the structuring element.If a pixel in the structuring element coincides with a foreground pixel in the image, then the input pixel is set to the foreground value and if the corresponding pixels in the image is in the background, then the input pixel is left as it is. Dilation is used for filling up of holes of certain shape and size which is given by the structuring element. The expression for Morphological Dilation is given by

$$A \oplus B = \{z | (\hat{B})_z \cap A \neq \Phi\}$$

- ***Opening:*** -Opening of an Image is nothing but combination of erosion and Dilation. Erosion operation is followed by Dilation operation using the same structuring element for both operations. Opening of an Image smoothens contours, break necks and eliminates protrusions. The expression of opening of an image is given by:-

$$A \circ B = (A - B) \oplus B$$

- ***Closing:*** **-**Closing of an Image is also a combination of Erosion and Dilation but the dual of Opening operation. Dilation operation is followed by Erosion Operation. Closing of an Image Smoothens Contours, Fuse narrow breaks and long thin gulfs, eliminate holes and fill gaps in contour. The expression of closing of an image is given by:-

$$A \bullet B = (A \oplus B) - B$$

After the morphological operation being performed on different frames, we would be left with the different objects as shown in figure

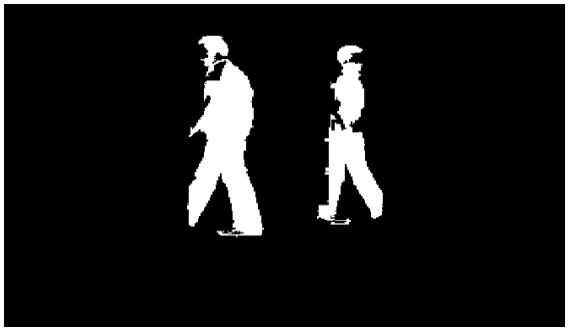
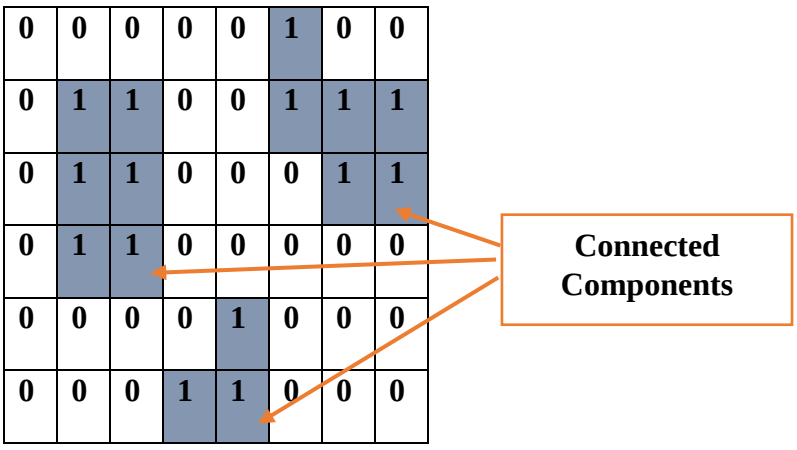


Fig. 3.12 Image after Morphological Operations

(b) ***Analysis of Moving Objects:*** **-**In an image, different moving objects are analysed by Connected Component Analysis. A set of pixels that form a connected group is known as a Connected Component in a Binary Image. This connected components detects connected regions in a binary image. They are mainly used for blob extraction on a binary image from a thresholding step. These blobs may be counted, filtered and tracked.



0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0

Fig. 3.13 Connected Components

The figure above shows the different Connected Components in an image. Connected Component Labelling is the process of identifying the connected components in an image by assigning each one a individual unique label. The figure is shown below:-

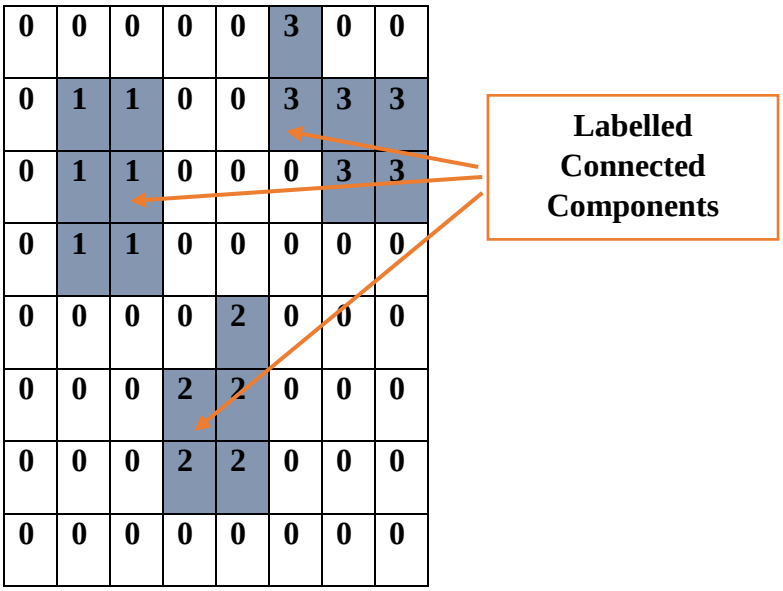


Fig. 3.14 Labelled Connected Components

bwconncomp computes connected components, as shown in the example:-

Fig. 3.15Binary Image after Morphological Operation

By using ***dd=bwconncomp(a,8)*** (matlab command of the connected component), where ***a*** is the input image and ***8*** is the connectivity of the pixels we would get the output as:-

Connectivity: 8; ImageSize: [288 384]; NumObjects: 4; PixelIdxList: {1x4 cell}

3.3.3Calculating the Properties of each Moving object: -The main aim is to detect the objects and decide whether the object is a human or a non-human. So, for this purpose, we need to find the properties of the objects which are present in the image. The properties of the image region is found by the ***region props***command in MATLAB. Now, by analysing the properties of this image we need to classify if the object is a human or a non-human.

3.3.4Classifying theObject: -After getting the number of objects and their properties, the classification has to be done about a human and a non-human. We know a moving human has a fixed w/l ratio. So, we have classified that the object is a human if the w/l ratio of a human lies between 0.30 and 0.75. So, the objects which falls in this range is classified as a human. The boundaries of each component is calculated and the width and length is calculated by:-

Wlratio = (ext1max - ext1min)/ (ext2max-ext2min)

Where *ext1max= right-most pixel of the object*

ext1min= left-most pixel of the object

ext2max= bottom-most pixels of the object

ext2min= top-most pixels of the object

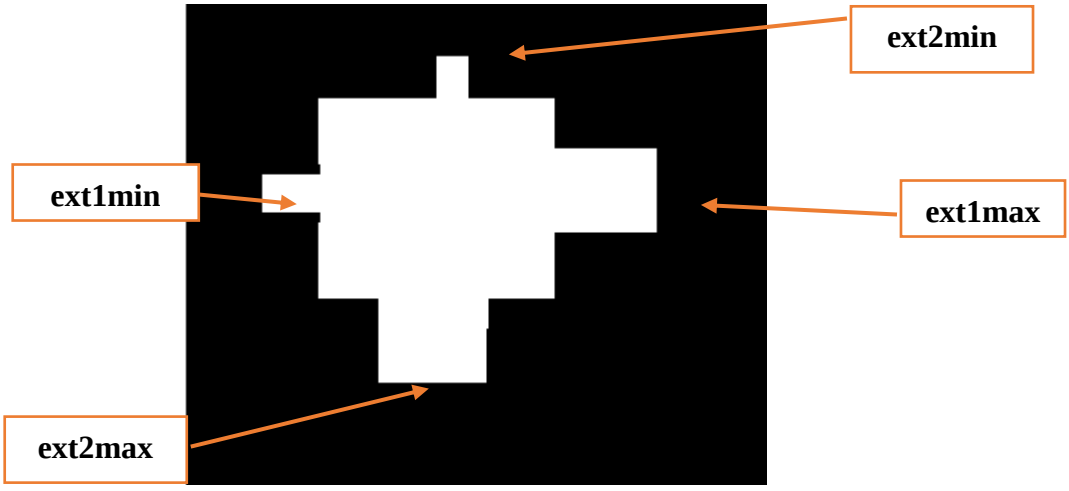


Fig. 3.16 Image showing the extreme points

The above figure shows the object with the extreme points which is calculated in the image. The wl ratio (aspect ratio) is calculated. If the wl ratio is found to be in the human range then a rectangular blob is constructed across the object to track the human.

REFERENCES

1) Yilmaz, A.; Javed, O. & Shah, M. (2006). Object tracking: a survey, ACM Computing Survey, Vol.38, No.13, pp.1-45.

2) Wren, C. R.; Azarbayejani, A.; Darrell, T. & Pentland, A. P. (1997). Pfinder: real-time tracking of the human body, IEEE Trans. Pattern Analysis. Machine Inteligent., Vol. 19, pp. 780–785.

3) McKenna, S.; Jabri, S.; Duric, Z.; Rosenfeld, A. & Wechsler, H. (2000). Tracking groups of people, Computer Vision: Image Understanding, Vol. 80, No. 1, pp. 42–56.

4) Paragios, N. & Deriche, R. (2000). Geodesic active contours and level sets for the detection and tracking of moving object, IEEE Trans. on Pattern Analysis and MachineIntelligent, pp. 266-280.

5) Schiele, B. (2000). Vodel-free tracking of cars and people based on color regions, Proceeding ofIEEE Int. Workshop Performance Evaluation of Tracking and Surveillance, Grenoble, France, pp. 61–71.

6) Coifman, B.; Beymer, D.; McLauchlan, P. & Malik, J. (1998). A real-time computer vision system for vehicle tracking and traffic surveillance, Transportation Research: Part C, Vol. 6, No. 4, pp. 271–288.

7) Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.

8) Koller, D.; Danilidis, K. & Nagel. H. (2000). Model-based object tracking in monocular image sequences of road traffic scenes, Int. Journal of Computer Vision, Vol.10, No.3, pp.257-281.

9) G. Welch and G. Bishop, "An Introduction to the Kalman Filter," UNCChapel Hill, TR95-041, July 24, 2006.

10) Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection." IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1063–6919, 2005

11) Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients", IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1491–1498.

12) Hui-Xing Jia and Yu-Jin Zhang. "Fast Human Detection by Boosting Histogram of Oriented Gradients". In ICIG 2007.

13) Li Zhihui, Shao Chunyan, Sun Di. "Real-time Human Detection Based on Cascade Frame". 2011 IEEE International Conference on Mechatronics and Automation, 514-518.

14) Xu Fen, Li Jie." Adaboost Human Detection based on a DSP Platform". IEEE International Conference on Electrical and Control Engineering, pages 335-338.

15) INRIA DATASET, *http://lear.inrialpes.fr/data.*

16) Changyan Li, Lijun Guo and Yichen Hu, “A New Method Combining HOG and Kalman Filter for Video-based Human Detection and Tracking,” In CISP, pp. 290–293, 2010.

17) Heewook Jung, Joo Kooi Tan, Seiji Ishikawa and Takashi Morie. "Applying HOG Feature to the Detection and Tracking of a Human on a Bicycle". 2011 11th IEEE International Conference on Control, Automation and Systems, pages 1740-1743

18) Li Li, XU Jining. "Moving Human Detection Algorithm Based on Gaussian Mixture Model". IEEE Chinese Control Conference, July 29-31, 2010, Beijing, China, pages 2853-2856

19) Xiaohui Liu, Zhigang Jin, Ming Gao, " A Robust Approach for Multi-Human Detection and Tracking", pages 832-835, 2012

20) David G Lowe. “Distinctive Image Features from Scale Invariant Interest Points,” International Journal of Computer Vision, pp. 91–110, 2004.

21) Schwartz W. S.,Kembhavi A.,Harwood D.,Davis L. S. "Human Detection Using Partial Least Squares Analysis". In Proceding of ICCV 2009.

22) R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. Lecture Notes in Computer Science, 3940:34–51, 2006.

23) Ya-Li Hou and Grantham K. H. Pang, "Human Detection In a Challenging Situation", 16th IEEE International Conference onImage Processing (ICIP), pages 2561-2564, 2009.

24) Stauffer,C.andGrimson,W.E.L“Adaptive background mixture models for real-timetracking”, IEEE Conference on Computer Vision and Pattern Recognition, 246-252, 1999.

25) Matsuyama, "The α-EMalgorithm: surrogate likelihood maximization using α-logarithmic information measures ", IEEE Transactions on Information Theory, 49(3):692-706, 2003.

26) PETS 2000 Dataset: http://ftp.pets.rdg.ac.uk/PETS2000

27) PETS 2000 Dataset: http://ftp.pets.rdg.ac.uk/PETS2000

28) Video Surveillance Online Repository, VISOR Dataset Collection. http://imagelab.ing.unimore.it/visor/

29) Songim Jia, "Robust Human Detecting and Tracking Using Varying Scale Template Matching ", IEEE International Conference on Information and Automation Shenyang, China, pages 25-30, June 2012.

30) P. Hidayatullah, H. Konik, "CAMSHIFT Improvement on Multi-Hue and Multi-Object Tracking", International Conference on Electrical Engineering and Informatics, July 2011.

31) Jialue Fan, Wei Xu, Ying Wu and Yihong Gong, "Human Tracking Using Convolutional Neural Networks", IEEE Trans. on Neural Networks, Vol.21, No.10, pp.1610-1623, October 2010.

32) Xiaofeng Lu, Li Song, Songyu Yu, Nam Ling, "Object Contour Tracking Using Multi-feature Fusion based Particle Filter", IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 237-242, 2012.

33) Tran Thi Trang , Cheolkeun Ha, "Irregular Moving Object Detecting and Tracking Based on Color and Shape in Real-time System", International Conference on Digital Object Identifier: PP. 415-419

34) Olga Zoidi, Anastasios Tefas, Ioannis Pitas: " Visual Object Tracking Based on Local Steering Kernels and Color Histograms". IEEE Trans. Circuits Syst. Video Techn. 23(5): 870-882 (2013).