Video Surveillance System for Security Applications

Vidya A.S.
Department of CSE
National Institute of Technology
Calicut, Kerala, India

V. K. Govindan
Department of CSE
National Institute of Technology
Calicut, Kerala, India

ABSTRACT
Computer Vision (CV) deals with replacement of human interpretation with computer based interpretation. It automatically analyses, reconstruct, and recognise objects in a scene from one or more images. Video surveillance is a topic in CV dealing with the monitoring of humans and their behaviours to analyze the habitual and unauthorized activities. An efficient video surveillance system detects moving foreground objects with lowest False Alarm Rates (FAR). This paper makes two proposals: one to detect the foreground in the video and the other to detect humans for surveillance applications. The proposed approach of foreground detection employs computations in YC_bC_r colour space for detecting moving objects in CCTVs. This system can handle slight camera movement and illumination changes. After foreground detection, the silhouette obtained is analysed and classified to determine whether it is humans or non-humans. In computer vision, usually human detection is based on human face, the head, and the entire body including legs as well as the human skin. In this work, the detection of humans is done based on the ratio of upper body and total height of silhouette. The precision and recall performance measures of the approach are computed and found to be superior to the existing Mixture of Gaussian approach in the literature.

General Terms
Computer Vision, Video surveillance.

Keywords
Computer vision, Video surveillance, Background modelling, Mixture of Gaussians, YC_bC_r color space, Human detection.

1. INTRODUCTION
Video surveillance systems are widely used in recent years in monitoring humans and their behaviours. This is used to analyze the habitual and unauthorized activities of humans. The applications of video surveillance include various areas of security systems, air port terminal check in, traffic monitoring, people counting etc. Recently, there is an increased research interest in this area for developing automatic surveillance systems for providing accurate surveillance. The motivation behind this work is due to the increasing pressures of guards in examining videos and the increasing cost of providing guarding services. Generally, a video surveillance system first constructs a background model using video frames having no moving objects. Based on this background model, incoming video frames are evaluated. A block diagram of general video surveillance system is given in Figure 1. This consists of four essential subsystems/ processes: a system for computing background model of the scene, image subtraction system for computing the motion of foreground image, separation of the silhouette of the foreground image, and a human detection system to provide whether it is a human or not.

Simple background subtraction approach is employed in [1, 2, 3] for the detection of motion. These provide incorrect motion information due to its sensitivity to illumination changes. For example, simple background subtraction method detects small movement of leaves of trees as foreground and can cause false alarms. Another drawback is that if a man parks a car, this should be included in the reference image, and subsequently this updated reference image should be used for background subtraction. That is, each time when there is change in the background, corresponding modification should be incorporated in the reference image. Thus this approach is not robust as one has to detect the changes in the environment manually.

Background modelling using Mixture of Gaussians (MOG) proposed by Stauffer and Grimson [4] solves many of the above problems. Here, each pixel is represented by a mixture of Gaussians. For example, one represents leaf of tree, another one for branch of a tree etc. An incoming frame is compared with the existing MOG model and a value that does not match with any one of the Gaussians at that position is detected as foreground.

Stauffer and Grimson [4] approach processed images in RGB color space. Processing an image in the RGB color space is not much efficient. For example, the modification of the intensity of a pixel involves the accessing of the three RGB values and performing the computation on all the three components. If the image is stored directly in the intensity and color format, some processing steps would be avoided. Many video standards use luminance and chrominance signals because of this reason.

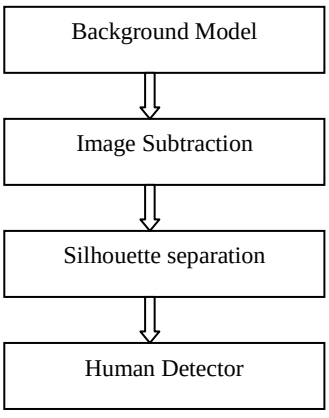

Fig 1: Block diagram of general video surveillance system

The rest of the paper is organized as follows: A literature survey on moving object detection and Human detection techniques is presented in Section 2. Section 3 deals with the proposed approach which involves processes for conversion

to YC_bC_r color space, foreground detection and Human detection. Results and their analysis are given in Section 4, and finally the paper is concluded in Section 5

2. RELATED WORKS

The problem of video surveillance involves two major sub-problems: one is the detection of moving objects and the separation of shadows and the other is the problem of determining whether the moving object is a human or not. So, this section briefly reviews some of the existing work in these two categories.

2.1 Moving Object Detection

Using image subtraction, moving foreground can be detected easily. Based on this simple approach, Rosin et al. [1] has proposed a work for detection and classification of intruders from the image sequence. The approach provides good sensitivity to changing pixels between successive frames, which is desirable feature for motion detection. However, the approach is sensitive to camera movement and illumination changes between adjacent frames.

Simple image subtraction will detect small movement in the background such as moving leaves of trees as moving object. Budi Sugandi et al. [2] proposed a method to reduce this kind of noise by using a low resolution image. The use of low resolution images is it eliminates the undesirable effect of the scattering noise and the small camera movement.

K. Srinivasan et al. [3] proposed a method for moving object detection employing an improved background subtraction from static camera video sequences. Here, first, the video data is converted into frames and these images are pre-processed to improve quality. Then, the background modelling creates an ideal background which should be static for all environment changes. Instead of normal background subtraction, here they used background subtraction using frame differencing and updated the threshold of each frame in a suitable manner using an automatic threshold updating algorithm.

Wren et al. [5] proposed a real-time system called *Pfinder* (person finder) for person segmentation, tracking, and interpretation. Maximum a posteriori probability approach is adopted by this method for the detection and tracking of humans. This system uses several domain-specific assumptions and when these assumptions break, the system the performance may degrade. Also, the approach cannot compensate for sudden changes in the scene. Another limitation is that it causes difficulty in segmentation of multiple persons in a scene.

In [6], Stauffer and Grimson proposed an approach based on mixture of Gaussian. They modelled the background as a mixture of Gaussians instead of modelling the values of all the pixels as one particular type of distribution. Pixel values that do not match with the existing Gaussians are grouped using connected components. These components are tracked using a multiple hypothesis tracker.

A technique for real-time surveillance of people and their activities is proposed by Haritaoglu et al. [7]. This method can be used for outdoor surveillance at night time and at low light level situations. In the first stage, foreground pixels are detected using statistical-background model and they are grouped into blobs. In the second stage, these blobs are classified into single person or multiple person or other objects. This system constructs appearance model for each person so that people can be identified even with occlusion. When light changes drastically, the performance of background subtraction will be poor. Also the system cannot segment shadows from the foreground.

The intelligent scene monitoring system performance limit estimation is studied in [8] by Sage and Wickham. They have investigated the performance limits of intelligent scene monitoring systems employing neural network classifiers in outdoor applications. The work defines the sterile zone analysis and image processing tasks, and presents some feasibility results.

The effect of colour space on tracking methods in video surveillance is demonstrated in the work [9] of Sebastian et al. The results obtained on different tracking methods demonstrate the superior performances for YCbCr and HSV color space than RGB color space.

2.2 Detection of Humans

There are different methods to detect human beings. Template based human detection method is based on predefined models defined by experts. Appearance based methods learn templates from a set of images or videos. Some of the examples of algorithms used by these approaches are briefly reviewed in the following:

In [10], Zhujie and Yu proposed a face recognition approach based on eigen faces. The first step is the construction of eigen faces of training images. A test image is recognized by computing the Euclidean distance and selecting the closest match. In [11], using *Neural networks*, certain features of human model such as area, perimeter, centroid, principal axis of inertia are selected and fed in to the classifier for training and classification. After exposure to different situations in the video, the model is updated with best possible match. In the approach of Osuna et al. [12] which employs support vector machines (SVM), the system works by scanning an image for face-like patterns at many possible scales and useing an SVM as its core classification algorithms to determine the appropriate class (face or non-face).

Detection based on skin is another method for human detection. Michael J. Jones et al. [13] proposed a method based on skin color. They construct color models for skin and non-skin classes from a dataset of nearly one billion labelled pixels. They compared performance of histogram and mixture models in skin detection and found histogram models to be superior in accuracy and computational cost.

A Bayesian approach based classification technique in YCbCr color domain for classifying skin color has been proposed by Chai et al. [14]. In this, the pixels are classified into skin color and non-skin color classes.

Zhengqiang Jiang et al. [15] proposed an approach for tracking pedestrians using smoothed colour histograms using an interacting multiple model framework. In this, pedestrians are detected in every video frame using the human detector technique proposed by Dalal and Triggs [16]. This method is more efficient than other tracking methods using Kalman filter [17] and colour histograms. [15] made use of Kernel Density Estimation (KDE) to get a more accurate and smoothed colour histogram.

The advantage of video is that humans can be detected based on the speed of motion in a sequence of frames. A model based vision for automatic interpretation of alarm has been proposed by Ellis et al. [18]. The model makes use of Hierarchical relationship to associate and match low level image data. Alarm causes are divided into two major classes-humans and non-humans. Each object model is partitioned into two components: the first part describes the individual instances of the animals, and the second describes the dynamic behaviour of the animals (speed, acceleration).

Human detection based on the ratio of height of the body to the height of the upper body part and ratio of height of the

body to the width of body is relatively easy. Employing the proposed approach of foreground detection, silhouette of moving object is isolated successfully. This silhouette is analyzed further based on the ratio. If the ratio falls within a given range it is most probably a human. This approach is made use of in the proposed work of this paper for human detection. For the detection of moving object, first a simple method is used to eliminate shadows to efficiently address the moving object detection.

3. PROPOSED APPROACH

In the proposed approach there are three major steps: 1) converting video into YC_bC_r color space; 2) foreground detection using Mixture of Gaussian method; and 3) techniques to determine whether the entered object is humans or non-humans. These are briefly presented in the following subsections.

3.1 The YC_bC_r Color Space

The YC_bC_r color consists of a luminance component (Y) and two chrominance components (Cb and Cr). Human eye is more sensitive to light changes compared to color changes. This color space makes use of this property. Here intensity component is stored with higher accuracy than the Cb and Cr components. In the original Mixture of Gaussian (MOG) model, Stauffer and Grimson [4] used the RGB components which are sensitive to illumination changes and shadowing, and are not independent. YC_bC_r color space is found to be more robust to these situations and have independent components of luminance and chrominance. Experimental results showed that YC_bC_r can handle shadows created by moving objects and can also reduce noise.

3.2 Foreground Detection

Mixture of Gaussian (MOG) is a widely used approach to detect moving objects from static cameras. Stauffer and Grimson [4] proposed a probabilistic background model, where each pixel is represented by a Mixture of Gaussians. First a background model is constructed based on 3-5 frames of video. A Gaussian mixture model can be formulated as [4]:

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \sum_{i,t}) \quad \textbf{(1)}$$

where X_t be the variable which represents the current pixel in frame I_t; K is the number of distributions; t represents time; $\omega_{i,t}$ is an estimate of the weight of the ith Gaussian in the mixture at time t; $\mu_{i,t}$ is the mean value of the ith Gaussian in the mixture at time t; $\sum_{i,t}$ is the covariance matrix of the ith Gaussian in the mixture at time t; η is a Gaussian probability density function.

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\Pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu)^T \Sigma^{-1} (X_t-\mu)} \quad \textbf{(2)}$$

For computational reasons, Stauffer and Grimson [4] assumed that the RGB color components are independent and have the same variances. So, the covariance matrix is of the form:

$$\sum_{i,t} = \sigma_{i,t}^2 I \quad \textbf{(3)}$$

where $\sigma_{i,t}^2$ is the variance of the ith Gaussian in the mixture at time t and $\sum_{i=1}^{k} \omega_{i,t} = 1$. Stauffer and Grimson [4] proposed to set K from 3 to 5. The initialization of the weight, the mean and the covariance matrix is made by using an EM algorithm. They used the K-mean algorithm for real time consideration. Every new pixel is checked against the existing K Gaussian distributions for a match. Based on the matching results, the background model is updated. X_t matches component i, if X_t is within 2.5 standard deviation of this distribution. Then, the weight is updated as

$$\omega_{i,t} = (1-\alpha)\omega_{i,t-1} + \alpha(M_{k,t}) \quad \textbf{(4)}$$

Where $M_{k,t}$ is 1 for the model which matched and 0 for the remaining models and α is the predefined learning parameter. The μ and σ parameters for unmatched distributions remain same. For matched distributions mean and variance are updated as:

$$\mu_{i,t} = (1-\rho)\mu_{i,t-1} + \rho X_t \quad \textbf{(5)}$$

$$\sigma_{i,t}^2 = (1-\rho)\sigma_{t-1}^2 + \rho (X_t - \mu_{i,t})^T (X_t - \mu_{i,t}) \quad \textbf{(6)}$$

$$\text{Where } \rho = \alpha P(X_t | \mu_{i,t-1}, \sum_{i,t-1}) \quad \textbf{(7)}$$

If none of the K distributions matches the current pixel value, the least probable distribution is replaced. New distribution has its mean value same as that of current value, an initially high variance, and low prior weight.

In the proposed method, first, the video is converted into YC_bC_r color space and the first 3-5 frames of video are used to construct a Mixture of Gaussian model. Then, the incoming frames are checked against this model and any changes in the environment will be updated. Processing video in YC_bC_r color space can effectively remove sudden intensity changes in the incoming frames before detecting foreground. This will reduce noise in the foreground detection.

After getting foreground image, morphological operations such as erosion and dilation are applied as postprocessing operations. These operations are done to remove misclassified and isolated pixels. This provides better results.

3.3 Human Detection in Video

The foreground image obtained is to be analyzed to determine whether it is human or non-human. If the obtained foreground has pixels greater than a particular threshold, then the corresponding frame of video is taken. The proposed approach first determines the upper body part of the foreground in the video frame using Viola Jones algorithm [19]. The height of this upper body part is calculated and total height of silhouette is also calculated. The ratio of the height of upper body and the total height of silhouette is taken to determine whether the object is a human or not. For human, this ratio will come within a particular range. Similarly ratio between width of the silhouette and total height is also taken to further enhance the accuracy of classification.

3.3.1 Upper body Detection

Upper body detection employs Viola Jones algorithm [19]. In this algorithm a set of features which are Haar Basis functions and more complex filters are used to detect different features of face and upper body. Figure 2 shows Haar-like functions used in this algorithm and how it is applied to an image. The sum of the pixels which are lying within the white rectangles are subtracted from the sum of pixels in the grey rectangles. In order to compute these Haar-like features in constant time, the integral image representation is introduced. Total number of Haar-like features is very large. In order to ensure fast classification, feature which are not helpful for discrimination between classes must be excluded and a small set of critical features must be used. For this feature selection, a modified version of AdaBoost is used. This method increases the speed of detection by combining successively more complex classifiers in a cascade structure.

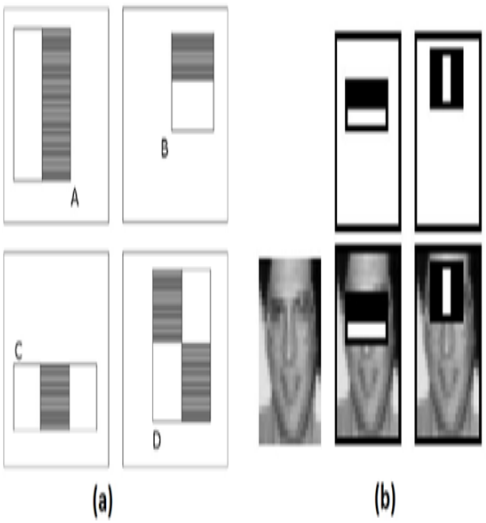

Fig 2: (a) shows different Haar-like functions used in Viola Jones algorithm [19] (b) shows applying Haar functions to an image

3.3.2 Body Parts Ratio Computation

From the above step, upper body part of foreground is detected. Then the height of upper body and total height of silhouette are determined, and the ratio between height of upper body and total height of silhouette is taken. For humans these ratio lies within the range of 0.2 - 0.3. Similarly the ratio of the width of the silhouette to the total height of the silhouette is computed. For humans, this ratio lies within 0.3-0.4. These two ratios are used to determine whether the detected object is human or non-human.

4. EXPERIMENTAL RESULTS

The system was implemented in MATLAB R2012a version 7.14.0.739. The entire motion detection process was tested over real videos. Time taken for a video of size 320 X 240 and having 351 frames is less than 15 seconds. Results were compared with the object detection technique given in [4] which uses MOG for background modelling in RGB color space. Experimental results showing the outputs of the proposed approach and the approach of Stauffer and Grimson [4] are given in Figures 3 and 4. The results demonstrate the superior visual performance of the proposed approach. This system converts video into different frames as in Figure 3(a). For constructing a Mixture of Gaussian model it uses the first 5 frames. After this, the incoming frames are compared based on this model. If any pixel value that does not fall into the Mixture of Gaussian model at that particular position is detected as foreground. The detected foreground based on Stauffer and Grimson [4] approach in RGB color space is given in Figure 3(b). The results demonstrate the inadequacy of the approach [4] to handle shadows and noises efficiently. The foreground detection based on YC_bC_r color space is given in Figure 3(c). The pictures in 3(c) clearly indicate that this system handles shadows and noise more efficiently. Also, when light changes, it will update the luminance component of the model so that noise due to light changes can be eliminated. The processing of videos in YC_bC_r color space is less time consuming than that of RGB color space.

Figure 5(a) shows upper body detection of a video frame containing foreground. For calculating the total height of silhouette efficiently, noises of area lesser than a particular threshold should be eliminated. Figure 5(b) shows silhouette after noise removal. Figure 6 shows upper body detection of non-humans. Table 1 shows ratio of height of upper body with that of total height of silhouette. This ratio falls within 0.2 to 0.3 for humans. Similarly ratio of width to total height of silhouette is also calculated. For humans this ratio falls within 0.3 to 0.4.

The performances of Mixture of Gaussian method [4] and the proposed foreground detection using MOG model in YC_bC_r color space are compared based on Precision and Recall measures:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \textbf{(8)}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \textbf{(9)}$$

Where TP is the True Positive, the case was positive and predicted positive; FN is the False Negative, the case was positive but predicted negative; and FP is the False Positive, the case was negative but predicted positive.

Figure 7 shows the comparison of the precision and recall performances of the proposed approach and the MOG approach in [4]. The proposed approach provides drastic improvements in precision which approaches about 0.96, near the ideal case of 1, whereas, in the Stauffer and Grimson approach [4], the precision is less than 0.92. The recall measure is also better than that of Stauffer and Grimson [4].

Fig 3: a) Original video frames

Fig 3: (b) Foreground detection by Stauffer and Grimson method [4]

Fig 3: (c) The proposed Foreground detection in YCbCr color space

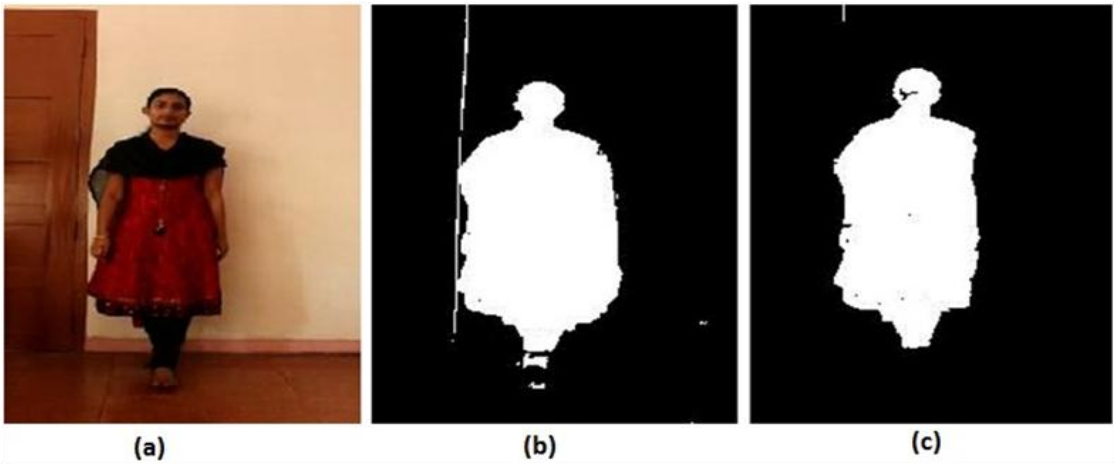

Fig 4: a) Original video frame, (b) foreground detection by Stauffer and Grimson method [4], (c) the proposed foreground detection in YCbCr color space.

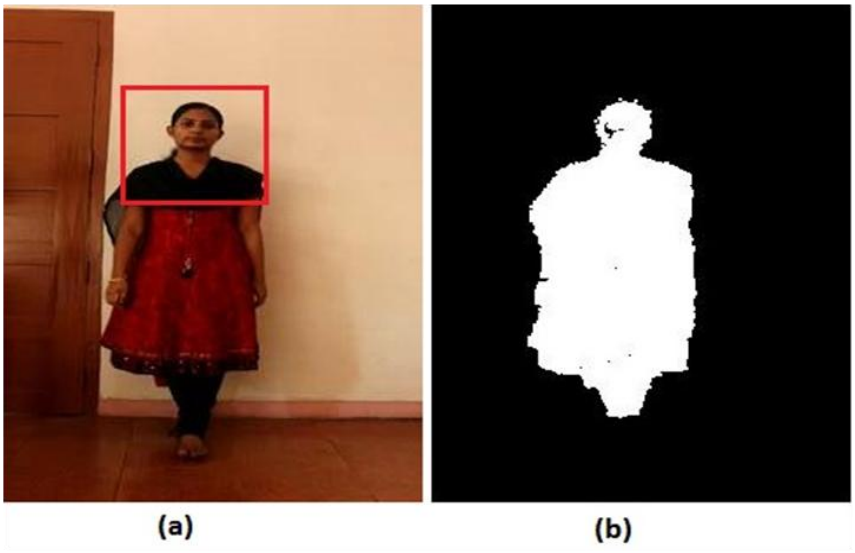

Fig 5: (a) Upper body detection in a video frame using Viola Jones Algorithm [19], (b) silhouette corrected after removing noises below a particular area.

Fig 6: Upper body detection of non-humans using Viola Jones Algorithm [19]

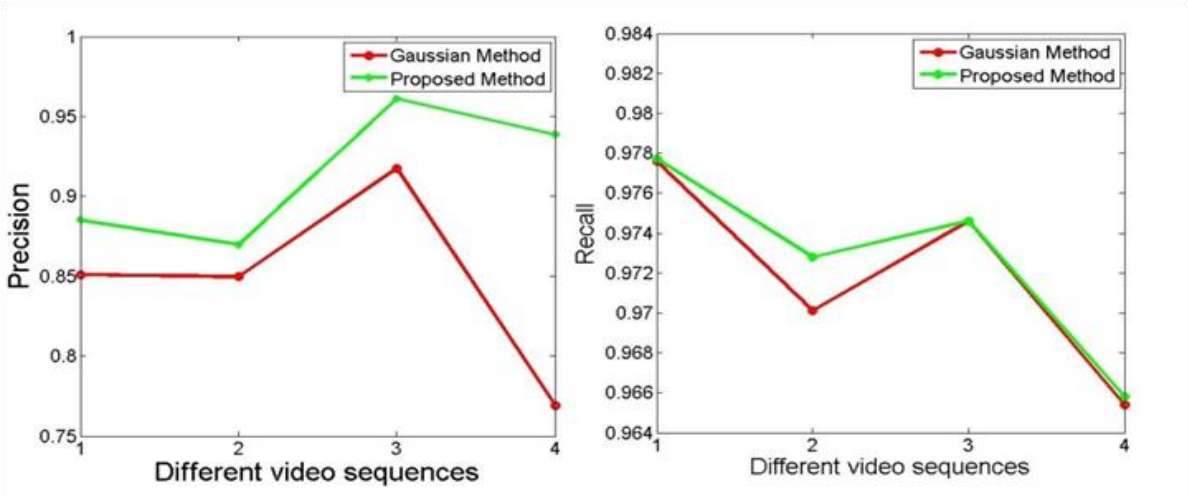

Fig 7: Graphs showing precision and recall of mixture of Gaussian (MOG) [4] approach, and the proposed approach for different video sequences

Table 1. Ratios of Upper body part to height and Width to height.

Category	Video frames	Ratio	
		Upper body part/Height	Width/Height
Humans	1	0.2727	0.3295
	2	0.2478	0.3943
	3	0.2392	0.3656
	4	0.2431	0.3942
	5	0.3365	0.4682
	6	0.2808	0.3314
Non-humans	7	0.5351	0.5718
	8	0.4121	0.7440
	9	0.5346	0.5623

5. CONCLUSION

Video surveillance is an active topic of current research. There remain many issues yet to be addressed. Back ground modeling by MOG can deal slow changes in light, long term scene changes, and swaying of trees. In this paper, an improvement for moving object detection employing the MOG model has been proposed. This is employed for human detection. By changing the color space to YC_bC_r, luminance and chrominance components are made independent and can handle the illumination changes. Also, this can provide better results if there are shadows created by moving objects and can reduce noise also. The performance comparison of the proposed approaches with the existing MOG approach demonstrates superior precision and comparable recall measures of the proposed approach. However, this approach suffers due to its sensitivity to sudden changes in lighting and hence the likelihood of causing false alarms. The moving object detected by the system is fed to human detector to classify the foreground object into humans and non-humans. It is observed that the system classifies humans and non-humans correctly based on the ratios of upper body part to height and the width to height of the moving object. However, upper body part detection may depend on the pose of foreground, that is, how it faces the camera. Occlusion of face by other objects affects upper body detection.

6. ACKNOWLEDGMENTS

The authors sincerely thank the members of faculty and staff of CSE Department, National Institute of Technology, Calicut for rendering all help and support for the successful completion of this work.

7. REFERENCES

[1] Rosin, P and Ellis, T, “Detecting and classifying intruders in image sequences”, in British Machine Vision Conference, 1991

[2] Sugandi, B. and Kim, H. and Tan, J.K. and Ishikawa, S.,”Tracking of moving objects by using a low resolution image”, Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference pp.408—408, 2007

[3] Srinivasan,K. and Porkumaran, K. and Sainarayanan,G., ”Improved background subtraction techniques for security in video applications” Anti-counterfeiting, Security, and Identification in Communication, 2009. ASID 2009. 3rd International Conference on pp.114--117, 2009

[4] Stauffer, C.; Grimson, W.E.L., "Adaptive background mixture models for real-time tracking," Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on , vol.2, no., pp.2 vol. (xxiii+637+663), 1999

[5] Wren, C. and Azarbayejani, A. and Darrell, T. and Pentland, A.,”Pfinder: real-time tracking of the human body” Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference pp.51—56,1996

[6] Stauffer, C. and Grimson, W.E.L.,”Learning patterns of activity using real-time tracking”,Pattern Analysis and Machine Intelligence, IEEE Transactions volume-22,number- 8 pp.747—757,2000

[7] Haritaoglu, I. and Harwood, D. and Davis, L.S, “W4: Real-time surveillance of people and their activities”, Pattern Analysis and Machine Intelligence, IEEE Transactions volume 22, number=8, pp.809—830, 2000

[8] K H Sage, K M Wickham, “Estimating performance limits for an intelligent scene monitoring system (ISM) as a perimeter intrusion detection system (PIDS)”, in International Carnahan Conference on Security Technology, 1994

[9] Sebastian, Patrick, Vooi Voon Yap, and Richard Comley. "Colour Space Effect on Tracking in Video Surveillance." International Journal on Electrical Engineering and Informatics 2.4 (2010): 306-320.

[10] Zhujie; Yu, Y.L.; , "Face recognition with eigenfaces," Industrial Technology, 1994., Proceedings of the IEEE International Conference on , vol., no., pp.434-438, 5-9 Dec 1994 doi: 10.1109/ICIT.1994.467155

[11] Ali, Syed Sohaib, M. F. Zafar, and Moeen Tayyab. "Detection and Recognition of Human in Videos Using Adaptive Method and Neural Net." In Soft Computing and Pattern Recognition, 2009. SOCPAR'09. International Conference of, pp. 604-609. IEEE, 2009.

[12] Osuna, Edgar, Robert Freund, and Federico Girosit. "Training support vector machines: an application to face detection." In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, pp. 130-136. IEEE, 1997.

[13] Jones, Michael J., and James M. Rehg. "Statistical color models with application to skin detection." Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.. Vol. 1. IEEE, 1999.

[14] Chai, D.; Bouzerdoum, A.; , "A Bayesian approach to skin color classification in YCbCr color space," TENCON 2000. Proceedings , vol.2, no., pp.421-424 vol.2, 2000

[15] Jiang, Z. and Huynh, DQ and Moran, W. and Challa, S,”Tracking pedestrians using smoothed colour

histograms in an interacting multiple model framework" Image Processing (ICIP), 2011 18th IEEE International Conference pp.2313—2316, 2011

[16] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893, June 2005.

[17] Z. Jiang, D. Q. Huynh,W. Moran, S. Challa, and N. Spadaccini, "Multiple pedestrian tracking using colour and motion models," Digital Image Computing: Techniques and Applications, pp. 328–334, Dec. 2010.

[18] Ellis, T.J, P.Golton,, "Model based vision for automatic alarm interpretation", in International Carnahan Conference on Security Technology 1990 pp. 62-67

[19] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57.2 (2004): 137-154.