A robust online motion detection and human identification system for indoor environments

Ahad Anwer, Tahir Q. Syed
National University of Computer and Emerging Sciences
Karachi 75300, Pakistan
ahad.anwer@gmail.com
syed.tahir@nu.edu.pk

Abstract:-Robust and efficient motion detection is one of the challenges in computer vision. It is a key technology to fight against terrorism, crime, public safety and for efficient management of traffic. Similarly significant is object classification, the process to classify moving objects into classes such as human or vehicle. In this paper, we have investigated the appropriateness of different background subtraction methods to overcome the problem of illumination variation and background clutter in an indoor environment. We present a technique for classification based on matching of head and shoulder silhouettes with a template learnt from training videos. The proposed method is found to be robust in both detecting motion in a video and deciding whether it is caused by a person.

Keywords:-Motion detection, object classification, background subtraction, comparisons of background subtraction techniques, human detection

I. INTRODUCTION

Motion detection and object recognition are regaining grounds as active research topics in computer vision [1, 3].They include problemssuch as to detect, classify and track objects over a sequence of images and together also makes the attempt to understand and describe object behaviour by replacing the traditional method of monitoring cameras by human operators.

Motion detection segments the moving foreground object from the rest image. Successful segmentation of foreground object helps in the subsequent process such as object classification, personal identification, object tracking and activity recognition in videos. Motion segmentation is done mainly with background subtraction, temporal differencing, and optical flow.we look at some of these methods later in the paper.

Detected moving foreground objects in an image sequence includes humans, vehicles and other moving object such as flying birds, moving clouds, animals, and abandoned object likes bags, luggage's, etc. It is necessary for a video surveillance system to classify the foreground objects into the different classes. The usual strategy, shape-based classification uses foreground objects area, apparent aspect ratio etc., as key features to classify into human, vehicle or any other moving object.

Visual surveillance systems[1] can provide effective and efficient application ranging from security. Surveillance applications are as follows:
Commercial and public security: Monitoring busy large places like market, bus stand, railway station, airports, important government buildings, banks for crime prevention and detection.
Military security: Surveillance in military headquarters, access control in some security sensitive places like military arms and ammunition store, patrolling of borders, important target detection in a war zone is done with surveillance systems.
Traffic surveillance: Monitoring congestion across the road, vehicle interaction, Detection of traffic rule violation [13] such as vehicle entry in no-entry zone, illegal U-turn can be done with visual surveillance systems.
Anomaly detection: Video surveillance system can analyze the behavior of people and determine whether these behaviors are normal or abnormal. Visual surveillance system set in parking area could analyze abnormal behaviors indicative of theft.

Our motivation in this work is to design a visual surveillance system for motion detection, and object classification. The objective is to develop an indoors system that could raise an alarm whenever it suspects a human walking through a restricted area. The applications of such a system would be in hospitals government and financial institutions' buildings and parking lots.

In this work we have tried to produce a simple, easy-to-reproduce indoor motion detection and object recognition system that could be used to signal out-of-hours activity or that in an out-of-bounds zone of the building. For this purpose, we have simulated various techniques available in the literature. A good background subtraction should be able to overcome the problem of varying illumination condition, background clutter, shadows, and camouflage and at the same time motion segmentation of foreground object should be done at the real time. It is hard to get all these problems solved in one background subtraction technique. So the idea was to simulate and evaluate their performance on videos created at our university campus, which is one of the places the system could be deployed. We have collected a dataset of video in a university hall for measuring the quality of different background subtraction technique used previously. We have collected a video sequence consists of 1210 frames of 320 × 240 resolution, acquired at a frame rate of 30 fps. In this video lightning conditions are good but there is a shadow castby some moving objects. The scene consists of a university hall, where two peopleare moving in and out from the video scene. Refer to the figure below for sample frames extracted for the purpose of illustration of the remainder of this paper.

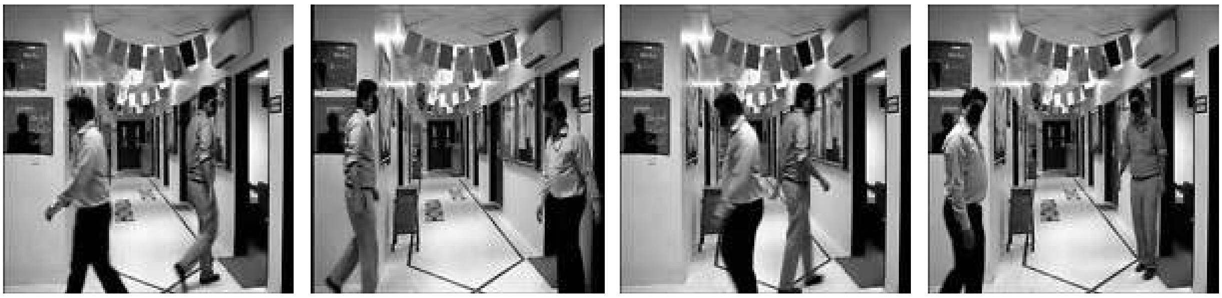


Figure 1: Frame exemplars used for explaining algorithmic results in later sections.

II. RELATED WORK

It is hard to get all above problems solved in one background subtraction technique. So the idea was to investigate different background subtraction techniques available in the literature and evaluate their segmentation quality on the video dataset we would be working with.

Background subtraction detects moving regions in an image by taking the difference between the current image and the reference background image captured from a static background during a period of time. The subtraction leaves only non-stationary or new objects, which include entire outline region of an object. The problem with background subtraction is to automatically update the background from the incoming video frame and it should be able to overcome the following problems:

Motion in the background: Non-stationary background regions, such as branches and leaves of trees, a flag waving in the wind, or flowing water, should be identified as part of the background.
Illumination changes: The background model should be able to adapt to gradual changes in illumination over a period of time.
Memory: The background module should not use much resource, in terms of computing power and memory.
Shadows: Shadows cast by moving object should be identified as part of the background and not foreground.
Camouflage: Moving object should be detected even if pixel characteristics are similar to those of the background.
Bootstrapping: The background model should be able to maintain background even in the absence of training background (absence of foreground object).

We now discuss some methods of background elimination or foreground extraction.

Image segmentation could be used for separating the background areas of the image from foreground regions of motion that are of interest discussed by Pal et al. [20]. There are some fundamental assumptions that the background will remain stationary. This assumption necessitates that the camera be fixed and that lighting not change suddenly.

Frame differencing is a pixel-wise differencing between two or three consecutive frames in an image sequence to detect regions corresponding to moving object such as human and vehicles discussed by M. Y. Siyal [18,19] . The threshold function determines change and it depends on the speed of object motion. It’s hard to maintain the quality of segmentation, if the speed of the object changes significantly. Frame differencing is very adaptive to dynamic environments, but very often holes are developed inside moving entities.

Barnichet al. [1] have discussed a simple methodology for background subtraction an absolute difference is taken between every current image and the reference background image to find out the motion detection mask. The reference background image is generally the first frame of a video, without containing foreground object.
If the absolute difference is greater than or equal to threshold value, the pixel is classified as foreground, otherwise the pixel is classified as background, and the threshold decides whether the pixel is a part of foreground or background.

In the previous algorithms the problem is that it also detects the motion in the background like (bushes, leaves, etc., noise in the camera and lighting). To overcome these problems dynamic learning of the background is important. The working for the algorithm is to capture N images and calculate the average of all images for modeling the background.
Cucchiara et al. [9] have presented a methodology in which we can use the *median* of the previous n frames as the background model. Assume that the background is more likely to appear in a scene.
The algorithm computes the median for each pixel (x, y) in the background image containing K frames.

$$\text{Background } B(x,y) = \text{Median}(l(x, y, t-i))$$

Sigari et al. [7] present the fastest and the most memory compact background modeling is *running average method* and also highlightits weaknesses. In this method, background extraction is done by arithmetic averaging on train sequence. Simple background subtraction cannot handle illumination variation and results in noise in the motion detection mask. The problem of noise can be overcome, if the background is made adaptive to temporal changes and updated in every frame. In this method, background is updated as follow:

$$B_t(x, y) = (1 - \alpha)\, B_{t-1}(x, y) + \alpha\, I_t(x, y)$$

Where the coefficient α represents the learning rate, a constant smoothing factor between 0 and 1 .A higher α discounts older observations faster.

Sigari et al. [7] present this methodology in which he have done some modification in the running average algorithm and the only difference is that the background image will only be updated if the pixel is not detected as moving. Modified running average method has better result with respect to standard running average method. Nevertheless, it has some drawbacks because of using hard limiter function in background subtraction and background updating.
The binary motion detection mask D(x, y) is calculated as follows: if the absolute difference is greater than or equal to threshold value, the pixel is classified as foreground, otherwise the pixel is classified as background, and the threshold decides whether the pixel is a part of foreground or background.

The problem with the previous algorithms are that the threshold is set to be same for all pixel, so the idea is to learn the background and its variations for each pixel. Grimson et al. [4] present the *adaptive background mixture model* approach in which a pixel is modeled by a mixture of K Gaussian distributions at time t. The probability of detecting the current pixel value is given by

$$P(X_t) = \sum_{i=1}^{k} \omega_{i,t} * n(X_t, \mu_{i,t}, \Sigma_{i,t})$$

where K is the number of distributions, $\omega_{i,t}$ is an estimate of the weight (what portion of the data is accounted for by this Gaussian) 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, $\Sigma_{i,t}$ is the covariance matrix of the ith Gaussian in the mixture at time t, and where η is a Gaussian probability density function.

$$n(X_t, \mu, \textstyle\sum i, t) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sum i, t|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu_t)^T \Sigma^{-1} (X_t - \mu_t)}$$

The algorithm could be summarized as follows:

```
For each time frame t do
        For each pixel (x, y) do
                For each Gaussian component i = 1 to K do
                        IF$|X - \mu_{k,t}| \leq 2.5 *\sigma_{k,t}$ then
                                $\omega_{k,t} = (1 - \alpha)\omega_{k,t-1} + \alpha(M_{k,t})$
                                $p = \alpha / \omega_{k,t-1}$
                                $\mu_{k,t} = (1 - p)\mu_{k,t-1} + p(X_t)$
                                $\alpha_{k,t} = (1 - p)\alpha^2_{k,t-1} + p(X_t - \mu_{k,t})^T (X_t - \mu_{k,t})$
                        Else
                                $\omega_{k,t} = (1 - \alpha)\omega_{k,t-1}$
                End
                Normalize weights $\omega_{k,t}$
                if none of the k distributions match, create new component
                Gaussians are ordered by the value of $\omega/\alpha$
        First B distributions are chosen as the background model $B = arg\,min_b\left(\sum_{k-1}^{b} \omega_k > T\right)$
    End
End
```

III. COMPARISON OF THE BACKGROUND MODELING TECHNIQUES

When we talk about segmentation quality evaluation there is no established standard procedure for the evaluation of its results, similar to the segmentation theory itself. A common classification of evaluation methods has been suggested by Zhang [10], classifying three alternatives: analytic methods, empirical goodness methods, and empirical discrepancy methods. In recent studies, empirical goodness methods are also referred to as unsupervised evaluation methods, empirical discrepancy methods are denoted as supervised or stand-alone evaluation methods e.g. Zhang et al. [11]. These evaluation techniques are also discussed and used by Vigneron et al. [12].

We use a method where the ground-truth is achieved by delimiting object boundaries by hand, and the resulting segmented image is used as reference image, using which different segmentations methods can be compared. What does not change is the nature of the segmentation method. The ground-truth serves to establish whether the pixel is in the correct basin or is elsewhere. Thus a group of pixels that should have been, according to the ground-truth, in a particular basin, may either be attributed by segmentation to another, or just as well the other way round. This gives rise to two types of errors in segmentation that the evaluation tries to quantify: Exceeding Pixels (EP) beyond object boundaries, resulting in unnecessary pixels recognized in object and Missed Pixels (MP) that is loss of object pixels to the background or neighboring basins. Exceeding pixel is calculated by measuring the ratio of segmented region pixels lying outside of the manual contour, and Missed pixels is calculated by measuring the ratio of the manually cut region pixels not in the segmented region.

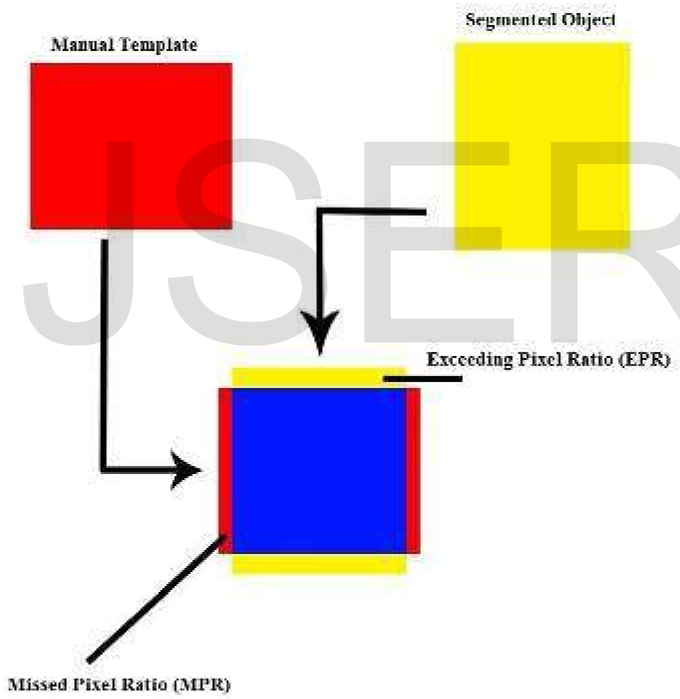


Figure 2: Illustration of the comparison [12].

A. Experimental Results

For measuring the quality of different background subtraction technique has been used previously. We have collected a video sequence consists of 1210 frames of 320 × 240 resolution, acquired at a frame rate of 30 fps. In this video lightning conditions are good but there is a shadow cast by some moving objects. The scene consists of a university hall, where two people are moving in and out from the video scene.

For measuring accuracy, different metrics such as *Missed Pixel Ratio* (***MPR***), *Exceeding Pixel Ratio* (***EPR***), and *Correct Pixel Ratio* (***CPR***) is calculated and tested with video sequence. The Correct pixel ration is calculated by taking the difference (MPR + EPR) by one.

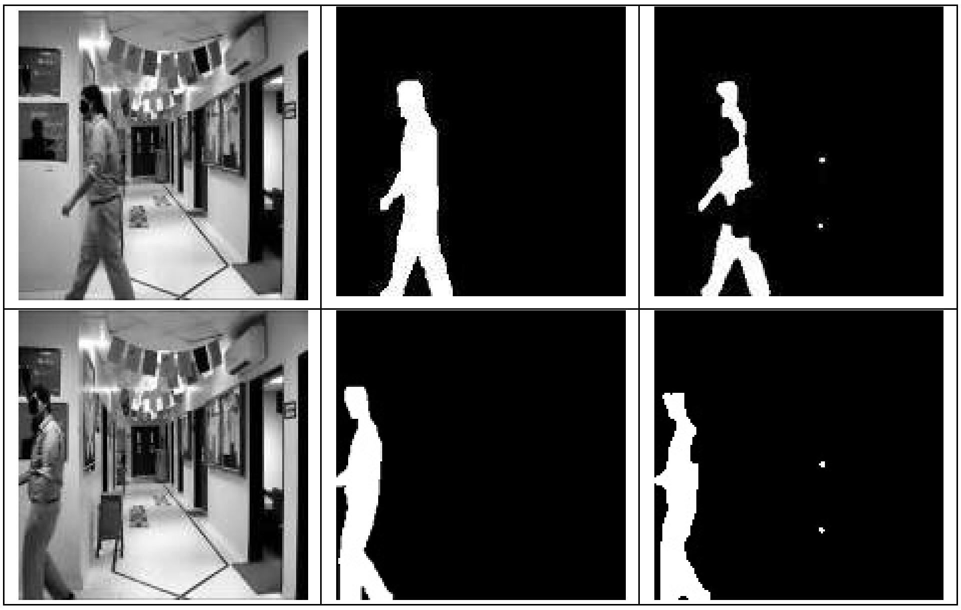
Correct Pixel Ratio (CPR) = 1 – (Missed Pixel Ratio (MPR) + Exceeding Pixel Ratio (EPR))

Table 2.2: Running average accuracy results

Fr. No.	CPR	MPR	EPR
1010	0.52853	0.47147	0.00011
1030	0.77499	0.17932	0.04568
1110	0.64716	0.30772	0.04511
1120	0.49071	0.46266	0.04662
1180	0.55733	0.55733	0.00420
1230	0.61644	0.61644	0.01511
1240	0.60989	0.38957	0.00053
Mean	0.60357	0.42635	0.02248
Accuracy	60.35%	42.26%	02.24 %

Table 2.3: Gaussian mixture model accuracy results

Fr. No.	CPR	MPR	EPR
1010	0.85279	0.13221	0.01499
1030	0.89160	0.06892	0.03947
1110	0.80492	0.07075	0.12432
1120	0.77876	0.10214	0.11910
1180	0.77558	0.16103	0.06339
1230	0.81705	0.15457	0.02838
1240	0.82183	0.16935	0.00881
Mean	0.82036	0.12271	0.05692
Accuracy	82.03 %	12.27 %	05.69 %



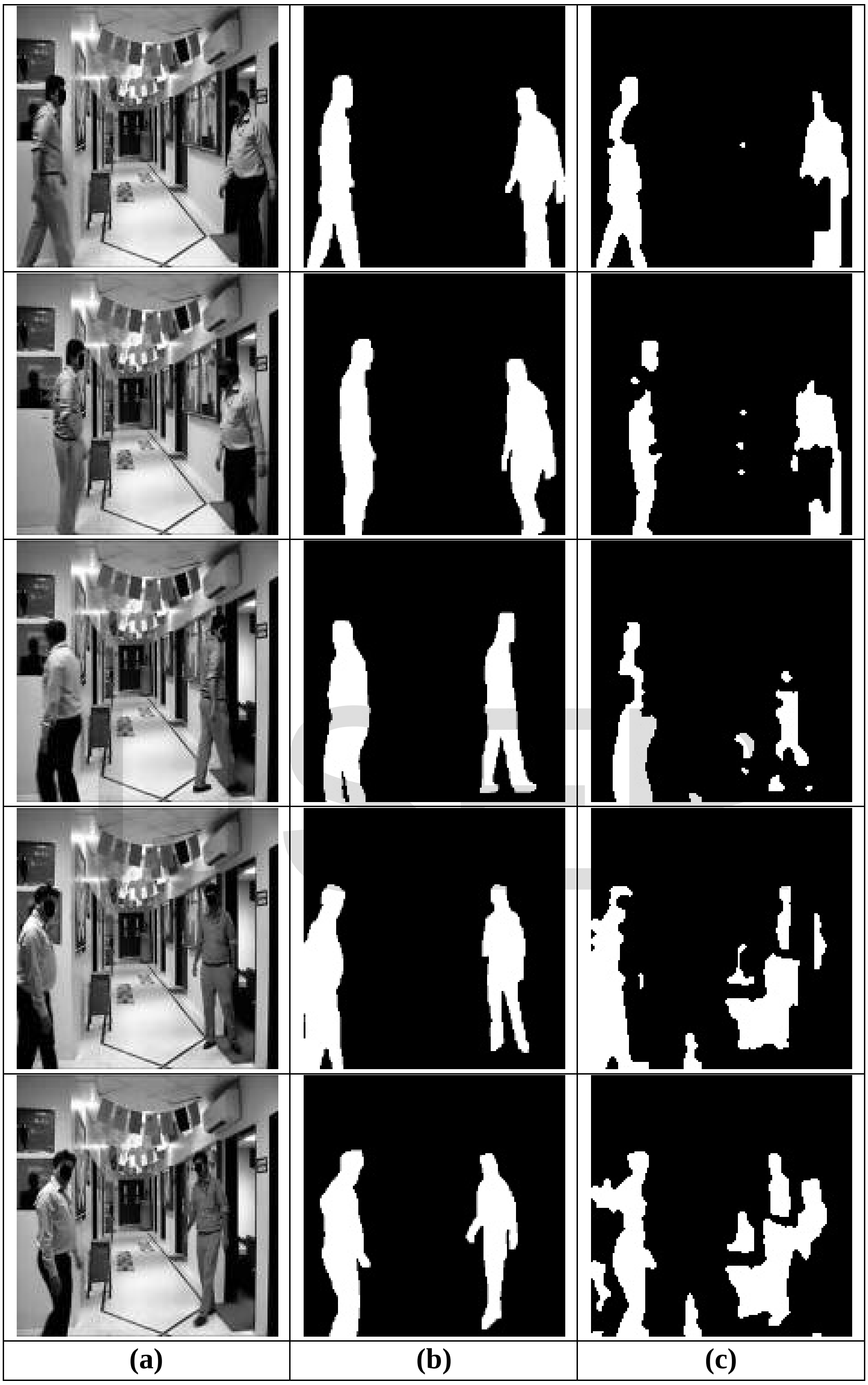
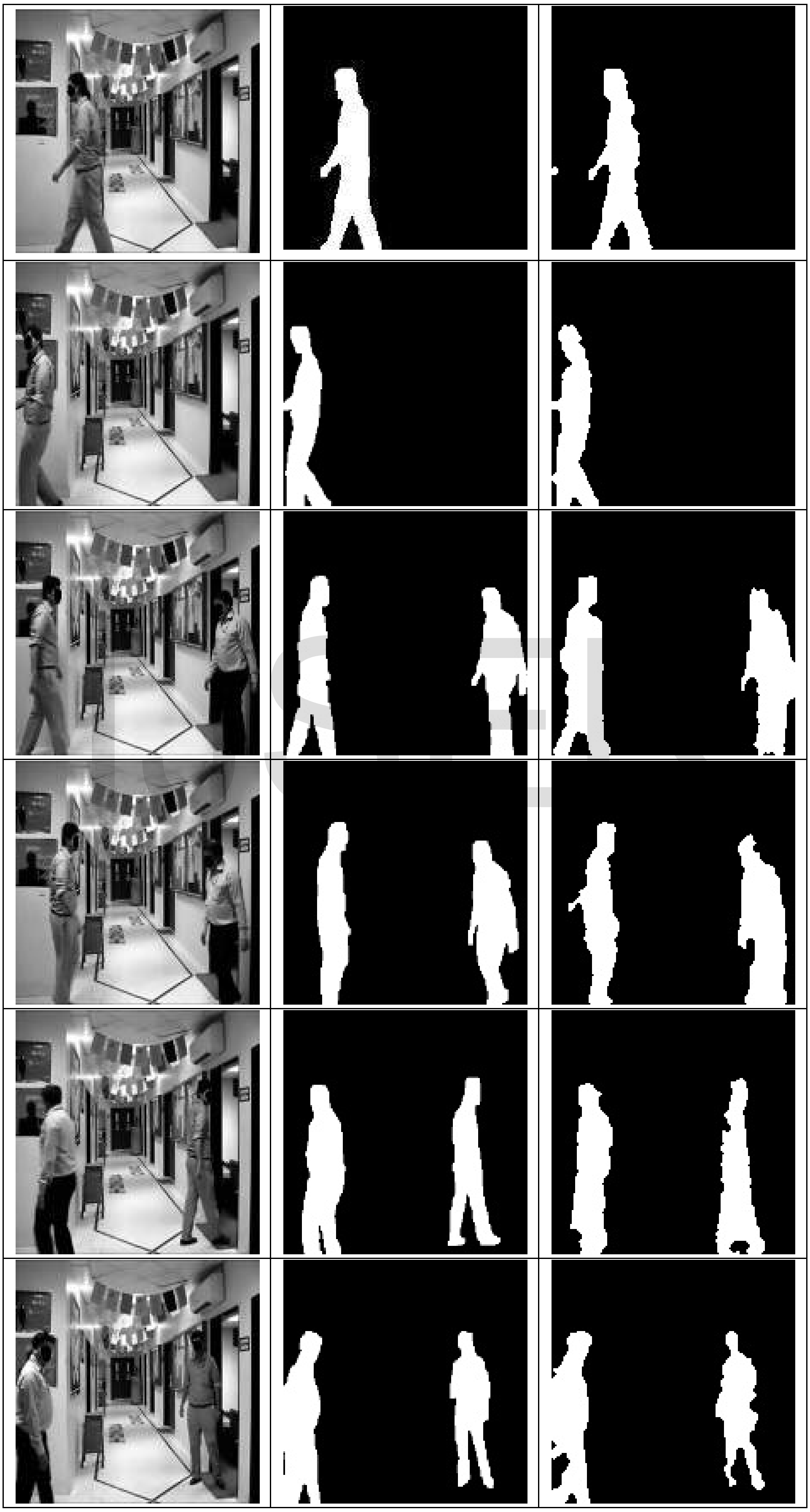
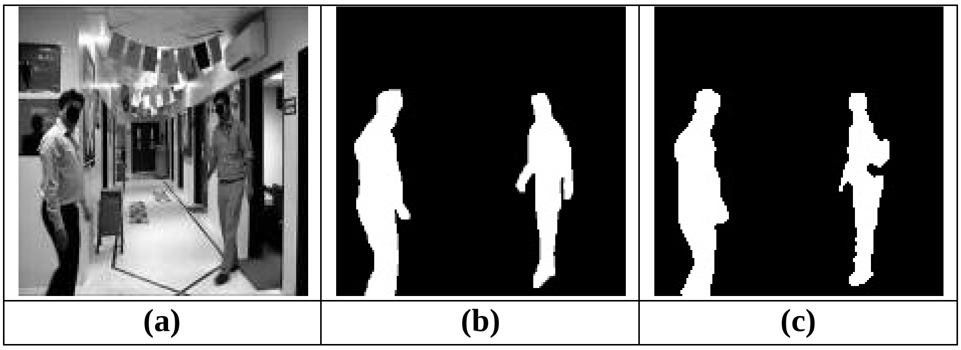


Figure 3: (a) Original image (b) ground truth (c) running average results

The above results shows the original intensity image (a), Manual segmented Ground truth images (b) and the results of the running average algorithm for different frames in a video sequences.





(a) (b) (c)

Figure 4: (a) Original image (b) ground truth (c) Gaussian MM results

The above results shows the original intensity image (a), Manual segmented Ground truth images (b) and the results of Gaussian Mixture Model algorithm for different frames in a video sequences.

The evaluation results have shown that the running average have scored an average CPR of 61.35 % although the MPR was 42.26% and the EPR was 02.24 %.The adaptive background mixture model is much more better than the previous algorithm and have scored an average CPR accuracy of around 82.03% with a decrease in the MPR to 12.27 % and a slight increase in the EPR to 05.69 %,so it has proven that *Adaptive Background Mixture Model* has results that are superior to the other algorithms and as an outcome of this comparison it seems logical to select the Adaptive Background Mixture Model for the rest of this work.

IV. OBJECT CLASSIFICATION

In visual surveillance, motion detection is the first important step that classifies the moving foreground object from the background. The segmented moving foreground object may be humans, vehicles, animals, flying birds, moving clouds, leaves of a tree, or any other noise etc. The job of a classification stage in the motion detection is to classify the moving foreground object into predefined classes such as single person, group of person, or vehicle, etc. The visual surveillance system is mostly used for humans and vehicles. Once the foreground object belongs to this class, the latter task such as personal identification, object tracking, and activity analysis of the detected foreground object can be done much more efficiently and accurately. The object classification is a standard pattern recognitionproblem and there are two approaches for classifying [3] moving foreground objects.

In shape-based classification, the moving foreground object region such as points, boxes, silhouettes and blobs are used for classification. Lipton et al. [15] have used image blob dispersedness and area to classify moving foreground object into human, vehicles and noise. If a target is present over a longer duration in the video, then the chances of having foreground object are high and if it is for short duration, then it is cluttered and is due to the noise in the frame. Dispersedness of an object is calculated from the given formulae.

$$\text{Dispersedness} = \frac{\text{Perimeter}^2}{\text{Area}}$$

Human body shape is complex in nature and will have more dispersedness than a vehicle. So the humans can be classified from vehicle using dispersedness and they have use Mahalanbois distance-based segmentation for foreground object classification. Rivlin et al. [16] have used a small set of features, like characterizing object shape and motion dynamics to classify objects.

The human body is non-rigid and articulated. It shows periodic motion and this property of human can be used for classification from the rest of foreground objects in the video frame. Cutler et al. [14] tracked interested object and its self-similarity is computed over time. For a periodic motion, computed self-similarity is also periodic. Time frequency analysis is done to detect and characterize the periodic motion.

In color-based classification, the moving foreground objects skin color is used for classification. Rehanullah [17] have stated that skin detection is a popular and useful technique for detecting and tracking humans and their body parts. The most attractive properties of color based skin detection are the potentially high processing speed and invariance against rotation, partial occlusion and pose change.

We have propose the *human silhouette based matching* methodology in which at very first the object to be classified is detected and segmented from video using Gaussian background subtraction technique which we have discussed above. In human silhouette template based classification the upper part which is head and shoulder of the foreground object, that is the top 30% of the foreground connected component) is compared with stored human silhouette template and serves to establish whether the pixel is in the correct basin or is elsewhere. Its similarity with the stored templates of the foreground object is found using the correct pixel ratio and is classified as human.

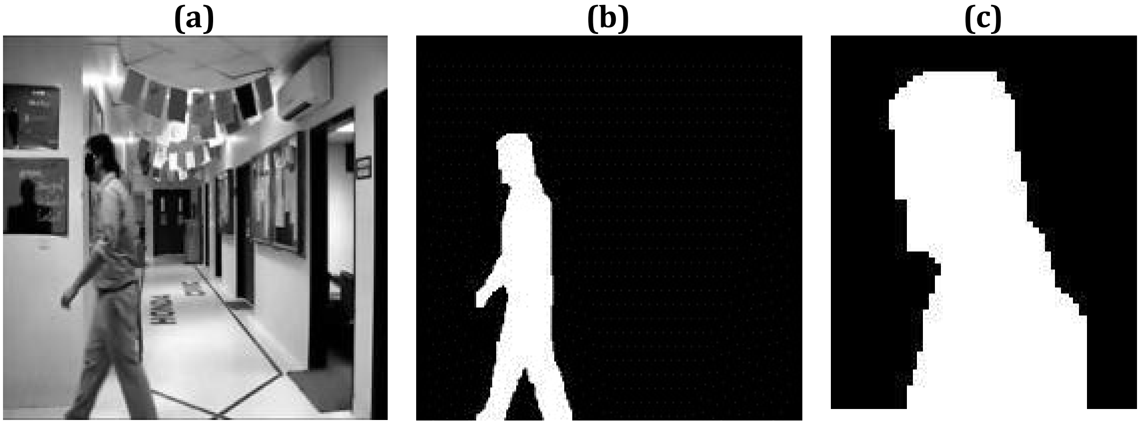


Figure 5: Illustration of object silhouette. (a) Original image (b) Segmented foreground(c) Object silhouette (head and shoulder).

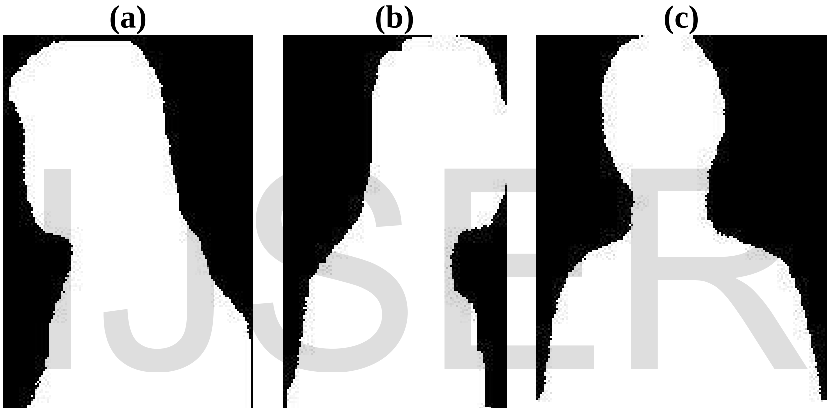


Figure 6: Stored template (head and shoulder) of the three major orientations of the upper human body

The template is built by averaging out the head and shoulder part of the connected component in motion in control videos, comprising 3 person profiles. The stored template is selected on the basis of movement made by an object. So if the object is moving from left to right form (b) is compared with the object silhouette. If it is moving from right to left form (a) is compared. If the object is moving from top to bottom or bottom to top from (c) is compared. So that we have to only compare the form which was selected with respect to the movement made by an object rather than comparing all forms.

The algorithm human silhouette template based classification only signals any object as human when the correct pixel ratio of the compared foreground object with the stored human silhouette template is greater than 60% to 70% percent, so that the object is classified as human.

V. CONCLUSIONS

In this work we have carried out work on motion detection in indoor videos, and on the classification of the object causing the motion as a human or not. In motion detection, we have studied different foreground extraction techniques available in the literature and have been implemented and evaluated their results. The selected method robustly extracts the foreground causing significant motion while discounting small motion as noise. We have also presented a solution for classifying detected objects into two groups: *human* and *non-human.* Future work includes the comparison of our proposed methodology with other techniques available in the literature for classifying objects to humans; and scaling the proposed methodology to include vehicles, animals and others.

REFERENCES

[1] O. Barnich and M. Van Droogenbroeck. Vibe: A Universal Background Subtraction Algorithm for Video Sequences. IEEE Transactions on Image Processing, June 2011.

[2] Piccardi, Massimo. "Background subtraction techniques: a review." InSystems, Man and Cybernetics, 2004 IEEE International Conference on, vol. 4, pp. 3099-3104. IEEE, 2004.

[3] Hu, Weiming, Tieniu Tan, Liang Wang, and Steve Maybank. "A survey on visual surveillance of object motion and behaviors." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 34, no. 3 (2004): 334-352.

[4] McIvor, V. Zang, R. Klette, The Background Subtraction Problem for Video Surveillance Systems, International Workshop Robot Vision 2001, Auckland, New Zealand, February 2001.

[5] Stauffer, Chris, and W. Eric L. Grimson. "Adaptive background mixture models for real-time tracking." In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 2. IEEE, 1999.

[6] Stringa, Elena. "Morphological change detection algorithms for surveillance applications." In Proc. British Machine Vision Conf, pp. 402-412. 2000.

[7] Sigari, MohamadHoseyn, NaserMozayani, and Hamid Reza Pourreza. "Fuzzy running average and fuzzy background subtraction: concepts and application."International Journal of Computer Science and Network Security 8, no. 2 (2008): 138-143.

[8] Christogiannopoulos, Georgios, Philip B. Birch, Rupert CD Young, and Christopher R. Chatwin. "Segmentation of moving objects from cluttered background scenes using a running average model." SPIE Journal 5822 (2005): 13-20.

[9] Cucchiara, R., Costantino G., M.Piccardi, and A. Prati. "Detecting moving objects, ghosts, and shadows in video streams." Pattern Analysis and Machine Intelligence, IEEE Transactions on 25, no. 10 (2003): 1337-1342.

[10] Zhang, Yu J. "A survey on evaluation methods for image segmentation." *Pattern recognition* 29, no. 8 (1996): 1335-1346.

[11] Zhang, Hui, Jason E. F., and Sally A. Goldman. "Image segmentation evaluation: A survey of unsupervised methods." *Computer Vision and Image Understanding* 110, no. 2 (2008): 260-280.

[12] Qasim,S. T. 2011. “Analysis of the migratory potential of cancerous cells by image preprocessing, segmentation and classification”. Universit´eD’evry Val D’ essonne. Ann Arbor: ProQuest/UMI.

[13] Buch, Norbert, Sergio A. Velastin, and James Orwell. "A review of computer vision techniques for the analysis of urban traffic." *Intelligent Transportation Systems, IEEE Transactions on* 12, no. 3 (2011): 920-939.

[14] Ross Cutler and Larry Davis. Robust real-time periodic motion detection, analysis, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22:781–796, 1999.

[15] Lipton, Alan J., Hironobu Fujiyoshi, and Raju S. Patil. "Moving target classification and tracking from real-time video." In *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*, pp. 8-14. IEEE, 1998.

[16] Rivlin, E., M. Rudzsky, R. Goldenberg, Uri Bogomolov, and S. Lepchev. "A real-time system for classification of moving objects." In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3, pp. 688-691. IEEE, 2002.

[17] Khan, R, Allan H., Julian Stöttinger, and Abdul Bais. "Color based skin classification." *Pattern Recognition Letters* 33, no. 2 (2012): 157-163.

[18] Fathy, M., and M. Y. Siyal. "Real-time image processing approach to measure traffic queue parameters." In *Vision, Image and Signal Processing, IEE Proceedings-*, vol. 142, no. 5, pp. 297-303. IET, 1995.

[19] Siyal, M. Yakoob, M. Fathy, and C. G. Darkin. "Image processing algorithms for detecting moving objects." In *International Conference on Automation, Robotics and Computer Vision (3rd: 1994: Singapore). Proceedings: ICARCV'94. Vol. 3*. 1994.

[20] Pal, N. R., and S. K. Pal. "A review on image segmentation techniques." *Pattern recognition* 26, no. 9 (1993): 1277-1294.