International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 2, April 2020, pp. 1337∼1345
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i2.pp1337-1345

Robust foreground modelling to segment and detect multiple moving objects in videos

Rahul M Patil, Chethan K P, Azra Nasreen, Shobha G
Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bangalore, Karnataka, India

Article Info

Article history:

Received Dec 9, 2018
Revised May 31, 2019
Accepted Oct 5, 2019

Keywords:

Background subtraction
Foreground modelling
Mean Averaging
Moving object detection
Video analysis

ABSTRACT

Last decade has witnessed an ever increasing number of video surveillance installations due to the rise of security concerns worldwide. With this comes the need for video analysis for fraud detection, crime investigation, traffic monitoring to name a few. For any kind of video analysis application, detection of moving objects in videos is a fundamental step. In this paper, an efficient foreground modelling method to segment multiple moving objects is implemented. Proposed method significantly reduces noise thereby accurately segmenting region of interest under dynamic conditions while handling occlusion to a large extent. Extensive performance analysis shows that the proposed method was found to give far better results when compared to the de facto standard as well as relatively new approaches used for moving object detection.

Corresponding Author:

Rahul M Patil,
Department of Computer Science and Engineering,
Rashtreeya Vidyalaya College of Engineering,
Bangalore-560069, Karnataka, India.
Email: patilmrahul06@gmail.com

1. INTRODUCTION

The first step in any video analytics solution is the segmentation of moving objects. Though this has been studied for several years, there has been lot of concerns when accurately detecting moving objects such as background noise, illumination changes, variable frame rate in recording videos resulting in lag, shadows and occlusion to name a few. In this paper, we propose an efficient object detection method that addresses issues such as background noise, illumination changes/reflection causing false positives, overlapping or occlusion to large extent, extracting exact bounding box or region of interest (ROI) using morphological operations and convex hull algorithm in post-processing phase. Various methods have been proposed for background subtraction [1,2], each having its own limitation due to many challenges such as sudden changes in scene, non-static background objects, lag introduced due to variable frame rate, changes in appearance of the objects with viewpoint and dynamic backgrounds such as gush of wind, movement of tree leaves, shadows etc. A review of the most relevant methods in background subtraction is provided in [3], giving a good understanding of the optimal method to be used for any background subtraction task. Segmentation methods using techniques such as background subtraction, Deep Learning etc., play highly pivotal roles in several applications, ranging from visual observation of animals [4,5] to video surveillance systems [6,7]. They are also extremely popular in content based video coding as in [8,9].

Much of the past and on-going research in this field aims at resolving these issues in order to improve accuracy of results [10]. Gaurav Takhar et al [11] discusses various methods of background subtraction such as basic, statistical as well as the machine learning techniques with the average, best and worst cases of several

other different methods. Proposed system is compared with statistical technique of adaptive Gaussian mixtures using popular datasets. Non-max suppression technique is discussed in [12]. A faster version of this method helps in the process of merging bounding boxes if multiple bounding boxes are obtained for a single object, which are in close proximity and have similar area sizes. For several morphological transformations that are used in the proposed method, sound understanding of these are provided in [13], most popular being Gaussian mixture model [14].

The state of the art in background subtraction has been proposed by [15], where an adaptive Gaussian mixture model is used to automatically find the number of Gaussian components for each pixel. A subsequent method is described in [16], where efficiency of the adaptive Gaussian mixture model is improved. Arun Varghese et al [1] discusses background subtraction being done at the pixel level and performance analysis using popular dataset Highway from `changedetection.net`. Performance analysis at the pixel level is also discussed in [17]. We used Pedestrians and Highway dataset from baseline category and Turnpike from the low frame rate category of the 2014 CDW datasets. Frame based performance metrics are discussed in [18,19] such as True Positives, False Positives, False Negatives and True Negatives for different datasets and models respectively.

The system proposed in this paper uses techniques such as fast non-maximum suppression method to increase the accuracy of detection, convex hull method to get better defined blobs of each foreground object and morphological transformations with circular kernels to get a much smoother outline of the detected foreground blobs. The model is extremely lightweight, very fast and requires no initial training. Proposed model also accounts for changing background by having the background updated by using weighted averages of each input frame. All in all, the model is computationally efficient, accurate for majority of the cases with a small number of limitations that will be discussed later.

2. GAUSSIAN MIXTURE MODEL

Pixels in the background are modelled with a mixture of K Gaussian distributions, the value of K being three to five. The time that a pixel stays in the scene is determined by the weights of the distributions in the mixture. The most likely background colours will be the ones that stay longer as determined by the weights. Improved Gaussian mixture model is more adaptive than the Gaussian mixture model [15,16], K distributions used for modelling is appropriately determined for each pixel in the image. The probability of a pixel having value X_N at time N is indicated in equation (1):

$$p(X_N) = \sum_{j=1}^{K} w_j \eta(X_N; \theta_j) \tag{1}$$

Wherein w_k is weight k^{th} Gaussian component. $\eta(x; \theta_k)$ is normal distribution of k^{th} component as indicated in equation (2):

$$\eta(x; \theta_k) = \eta(x; \mu_k, \sum_k) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \sum_k^{-1} (x-\mu_k)} \tag{2}$$

In which mean is μ_k and the covariance is $\sum_k = \sigma_k^2 I$. The K distributions are sorted based on the value of w_k / σ_k and the first B distributions are used to create a model of the background of the scene. B is computed as in equation (3):

$$B = \arg\min_b \left(\sum_{j=1}^{b} w_j > T \right) \tag{3}$$

Where T is the minimum fraction of the background model. In other words, it is the minimum prior probability that the background is in the scene.

(a) GMM adaptive to variable lighting conditions: This method incorporates per pixel Bayesian segmentation into the Gaussian mixture model in order to account for videos recorded in variable lighting conditions [20].

(b) Adaptive variable frame rate coding: This method adjusts the frame-rate of the video dynamically and adaptively, making use of information from already existing video encoders [21].

(c) Intermittent motion coding: This method involves disabling of motion coding during periods of inactivity in the video. Thus it records only parts of the video were active foreground movement is involved for further processing [22].

All of the methods explained above incur considerable overhead with regard to time or CPU usage. The Gaussian mixture model based methods cannot efficiently deal with variable frame rates in videos. The variable frame rate coding techniques make use of video encoder information, the compilation of which involves CPU overhead. Also, recording only during periods of activity means that the definition of activity in the scene has to be pre-determined in advance, and done so using extensive statistical analysis. Non-static background objects must be included the background modelled.

3. PROPOSED SYSTEM

Background is modelled by obtaining the background scene without occurrence of any of the foreground objects, so that foreground objects from it can be obtained by background subtraction. Though it looks simple, it is very difficult and a tedious task as it should not contain any foreground objects in it, i.e any movement such as gush of a wind, movement of tree leaves etc. should be part of the background itself. The background of the scene should be updated as and when the scene changes and must be free from any kind of noise and must be susceptible to any kind of illumination changes.

3.1. Running average method

A background model has to be constructed initially in order to perform the background subtraction task. Running average is found to be a good method of approximating the background. This method is faster than Gaussian mixture model and is more consistent than direct frame differencing [23]. Proposed system uses fast running average method for background modelling as illustrated in equation (Eq. 4):

$$dst(x,y) = (1-r).dst(x,y) + r.src(x,y) \tag{4}$$

Where $dst(x,y)$ is the accumulator image with the same number of channels as input image, $src(x,y)$ is input image which can have 1 or 3-channels, and r is a weight of the input image. Using continuous frames in a video stream, the weighted average background model can be calculated by choosing an appropriate value for r, for that particular sequence. By using a higher value of r, we are able to eliminate the foreground objects that are not persistent in the scene. Also, a suitable value of r can be chosen by taking into consideration the amount of data available for modelling. The process of learning the background is as illustrated in Figure 1.

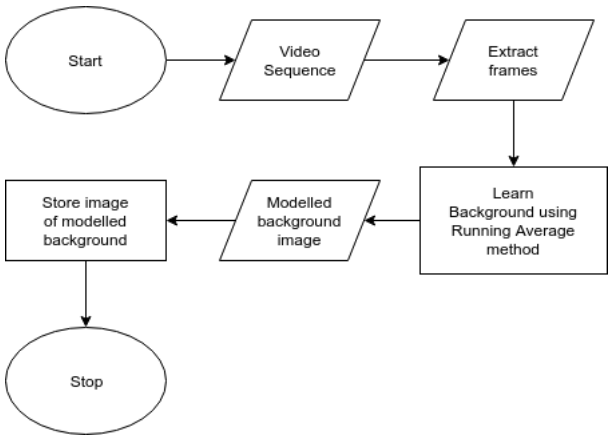


Figure 1. Running Average to learn background

3.2. Background subtraction

The V channel of the HSV image is fed as an input to the differencing method, where the absolute difference between the V channel of the current frame and the modelled background is obtained. This is done by finding the absolute difference between each pixel element of the modelled background and the V channel

of the current frame, which are fed as parameters to the method. The HSV color space is used because it works well against shadows [24]. The final absolute differenced image is processed to find and draw the most prominent contours for the detected foreground objects.

Then a thresholding is performed where pixels below a certain threshold value are assigned a 0 value, and the pixels having a value greater are assigned the maximum value of 255. This method is known as binary thresholding as shown below:

$$dst(x,y) \leftarrow \begin{cases} maxVal & \text{if, } src(x,y) > thresh \\ 0 & otherwise \end{cases} \tag{5}$$

Here $src(x,y)$ is a source image pixel, $thresh$ is the threshold value used in binary thresholding and $dst(x,y)$ is the result image pixel. $maxVal$ is the value that the particular $src(x,y)$ pixel will obtain if it's value exceeds that of the pre-assigned $thresh$ value. The entirety of the steps performed in the proposed method can be expressed in a flow diagram as seen in Figure 2. The Sequence of operation are shown in Figure 3.

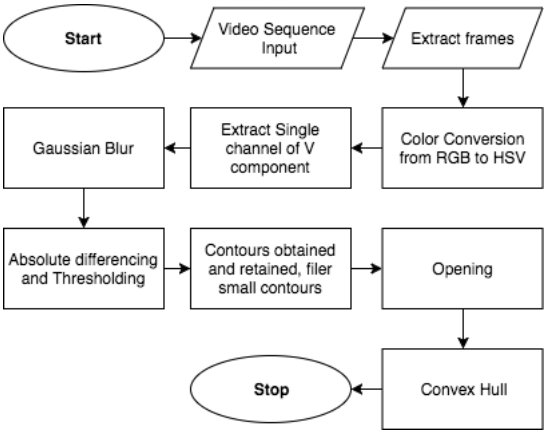


Figure 2. Proposed method to segment moving objects

(a) Difference Image (b) Gaussian Blur Applied (c) Thresholding, removing noisy contours and opening operation (d) Final contours obtained after Convex Hull

Figure 3. Sequence of operations

3.3. Foreground modelling

After the threshold frame is determined, we have a binary frame with blobs representing foreground objects. Morphological transformations such as dilation, erosion and opening are applied to reduce merging of contours of different foreground objects. Opening operation is used to eliminate portions of the foreground object that may just extend out into the background. It is achieved by using the dilation and erosion operation which augments and shrinks a region respectively. We use a structuring element S otherwise known as a kernel to perform these operations. This operation is used to expand the foreground object's obtained contours. The dilation of an image B with S, is given by the below equation (6):

$$B \oplus S = \bigcup_{b \epsilon B} S_b \tag{6}$$

Erosion reduces the size of the foreground object's contour and is used to remove unwanted excess contour elements that may have extended into the background. Similar to dilation, the erosion of an image B with structuring element is given below:

$$B \ominus S = \{b \,|\, b + s \,\epsilon\, B \,\forall\, s \,\epsilon\, S\} \tag{7}$$

Opening operation is erosion operation followed by dilation operation, which is used to prevent merging of contours of different objects and ultimately gives much better final bounding boxes for the foreground objects, and can be represented mathematically as in equation (8):

$$B \circ S = (B \ominus S) \oplus S \tag{8}$$

These blobs are extracted as contours. Smaller blobs and contours that represent noise and other unwanted detail are eliminated and properties like the edges, centres and areas of the final set of resulting contours are calculated. A convex hull of the contours is found to give a definitive shape to any incomplete contours that might have resulted due to similarity of intensity value or illumination defects.

In order to get whole bounding boxes for foreground objects, there was a need to make the contours of the foreground objects more wholesome. To accomplish this, the convex hull operation is performed on the contours. The contours obtained finally after performing this are used to draw the bounding boxes for the detected foreground objects. The convex hull of a finite set of points S is the set of all convex combinations of the points. Each and every point in this set denoted by x_i is attributed with a weight α_i. Each and every weight must be non-negative and their sum must be equal to unity. These weights are used to obtain a weighted average of all the points in set S. For various choice of coefficients, a certain convex combination is obtained that is a point in the convex hull. Therefore, the entire convex hull may be obtained by considering all the various combinations of weights. It can be expressed in a single equation as shown below in equation (9):

$$Conv(S) = \sum_{i=1}^{|S|} \left\langle \alpha_i x_j \mid (\nabla i : \alpha_i \geq 0) \wedge \sum_{i=1}^{|S|} \alpha_i = 1 \right\rangle \tag{9}$$

The final blobs are returned as contours, and the bounding boxes for all these contours are obtained and stored in an array structure. Then redundant bounding boxes that occur inside other larger bounding boxes are eliminated. Finally an iteration of fast non-max-suppression is employed to merge multiple detections for the same object for improved final results. It uses area of the obtained boxes in addition to the overlapping percentage of neighbouring boxes. Then the final boxes that are in the array are drawn onto the frames. Area of these bounding boxes along with their pixels are compared with the bounding boxes and the pixels of the ground truth frames in order to estimate and analyse the performance.

4. EXPERIMENTAL SETUP AND RESULT ANALYSIS

Dataset used for performance evaluation is CDnet, (Change Detection), consists of 31 videos depicting indoor and outdoor scenes with boats, cars, trucks, and pedestrians that have been captured in different scenarios and contain a range of challenges. Pedestrians and Highway from baseline category and Turnpike from the low frame rate category of the 2014 CDW datasets have been used. The validation metrics that have been used in the context of comparing the segmented result with the corresponding ground-truth for that frame in the video sequence are:

(a) **True Negative (TN)**: Pixels correctly classified as the background

(b) **True Positive (TP)**: Pixels correctly classified as the foreground

(c) **False Positive (FP)**: Pixels wrongly classified as the foreground

(d) **False Negative (FN)**: Pixels wrongly classified as the background

Various performance metrics that have been used are as shown from equation (10) to (17) below:

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

$$Recall(R) = \frac{TP}{FN + TP} \tag{11}$$

$$Specificity = \frac{TN}{FP + TN} \tag{12}$$

$$False\,Negative\,Rate = \frac{FN}{FN + TP} \tag{13}$$

$$False\,Positive\,Rate = \frac{FP}{FP + TN} \tag{14}$$

$$PWC = \frac{FP + FN}{TN + TP + FP + FN} * 100 \tag{15}$$

$$F - Measure = \frac{2RP}{R + P} \tag{16}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{17}$$

Table 1 shows the performance comparison of the proposed system, against the improved adaptive Gaussian mixture model [15] on three datasets, namely highway, turnpike and pedestrians.

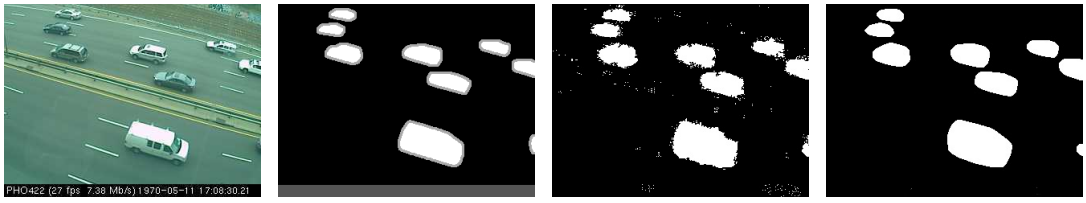
Table 1. Performance evaluation of proposed system with improved adaptive Gaussian mixture model and Hybrid model

Datasets		**Highway**			**Pedestrians**			**Turnpike**
Model	**Proposed**	**Zivkovic[15]**	**Hybrid**	**Proposed**	**Zivkovic**	**Hybrid**	**Proposed**	**Zivkovic**
Recall	0.7387	0.9619	0.9152	0.6594	0.9860	0.7290	0.9259	0.9649
Specificity	0.9982	0.9272	0.9314	0.9988	0.9613	0.9921	0.9868	0.9695
FPR	0.0137	0.5682	0.5391	0.0216	0.6804	0.1384	0.0724	0.1678
FNR	0.0334	0.0049	0.0118	0.0194	0.0008	0.0154	0.0134	0.0064
PWC	3.1237	6.8897	7.0895	1.9496	3.7379	2.2092	2.2525	3.1197
Precision	0.9817	0.6286	0.6293	0.9682	0.5917	0.8404	0.9275	0.8519
F-Measure	0.8430	0.7603	0.7453	0.7845	0.7396	0.781	0.9267	0.9049
Accuracy	0.9688	0.9311	0.9288	0.9792	0.6258	0.9767	0.9775	0.9688

As indicated in Table 1 it was found that proposed method was found to be effective and yielded better accuracy of 96.88% and precision of 98.17%. Also, it has a very low false positive rate and false negative rate for detecting moving objects in videos, when compared to the de facto standard of the improved adaptive Gaussian mixture model on the highway dataset from change detection net. The snapshots obtained with proposed system, and the adaptive Gaussian mixture model for three datasets, as shown in Figures 4, 5 and 6.

Table 1 also shows comparison of the proposed system with another existing method, namely the multi-modal hybrid approach of adaptive Gaussian mixture model and mean averaging. The hybrid model used for comparison can model and track moving objects in a video and it works as follows. In order to smoothen the extracted frames, a sequence of smoothing filters are applied, these being Gaussian blur and median blur, respectively. The approach taken to reduce noise uses the morphological operations erosion and dilation. Mean averaging is used for background modelling and frame differencing along with the adaptive Gaussian mixture model is used to obtain foreground masks. Contours are found from the foreground masks on which convex hull is applied to get the final object blobs. Proposed hybrid model is able to detect and track moving objects in videos in real time and is tested for many outdoor scenes, and snapshots of the obtained results follow the conclusion section. No comparison has been made for the Turnpike dataset for the hybrid model, as it has not been designed for low frame rate videos, and therefore it has not been included in the table.

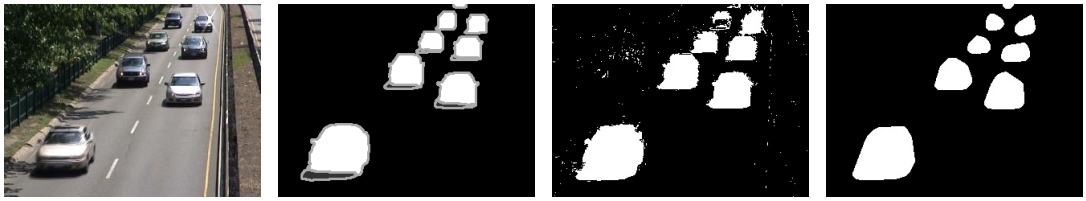
As evident from the Table 1, the proposed system is able to perform well when compared to the hybrid method as well. Effectively reduces noise and is able to segment exact ROI of moving objects. This is achieved by Gaussian blur and removal of small contours leading to noise, and by applying opening morphological operations. This isolates contours of different bounding boxes, even if the distance between the objects is small, thereby handling occlusion to an extent.



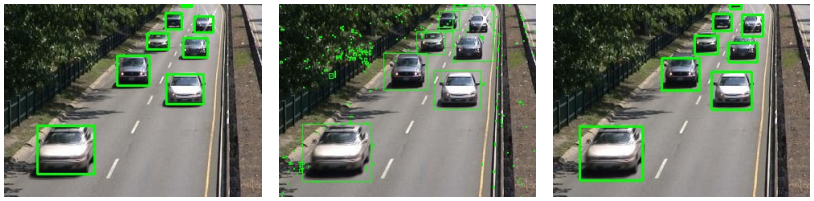
(a) Original Input (b) Ground Truth (c) Contours obtained by MoG2 (d) Contours obtained by Proposed Model

(e) Bounding boxes from Ground Truth (f) Bounding boxes from MoG2 (g) Bounding boxes from Proposed Model

Figure 4. Results for Turnpike dataset

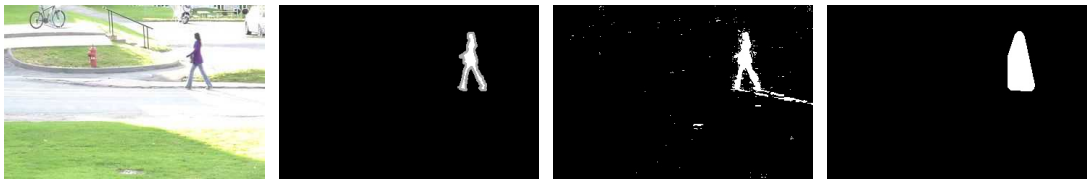


(a) Original Input (b) Ground Truth (c) Contours obtained by MoG2 (d) Contours obtained by Proposed Model

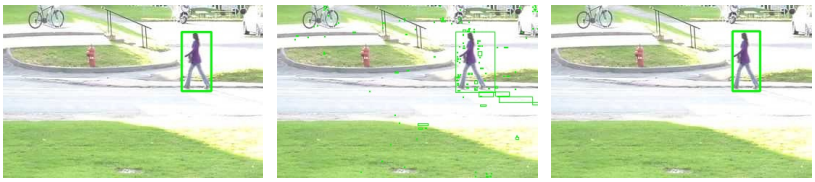


(e) Bounding boxes from Ground Truth (f) Bounding boxes from MoG2 (g) Bounding boxes from Proposed Model

Figure 5. Results for Turnpike dataset



(a) Original Input (b) Ground Truth (c) Contours obtained by MoG2 (d) Contours obtained by Proposed Model



(e) Bounding boxes from Ground Truth (f) Bounding boxes from MoG2 (g) Bounding boxes from Proposed Model

Figure 6. Results for Turnpike dataset

The Figures 4, 5 and 6 show a comparison of the working of our proposed model against Improved Adaptive Gaussian mixture model and the Hybrid model. Each figure consists of a set of 7 sub-figures each, which summarize the results obtained on the different datasets that have been used. The first sub-figure, is the input frame from the original dataset, just as is the following ground truth sub-figure. The following two sub-figures are the blobs that are obtained by the improved adaptive Gaussian mixture model and our own method respectively. The following three figures are as their captions suggest. Essentially, they are bounding boxes that have been obtained for the corresponding blobs, and drawn onto the original input frame.

5. CONCLUSION

The proposed system was found to be an effective approach in capturing small and large movements in the moving objects and extracts well defined foreground objects. Exact region of interest were extracted and it yielded better accuracy when compared to state of art development method such as mixture of Gaussians and relatively new hybrid approach of mean averaging and mixture of Gaussians method when it comes to issues such as noise and much better contours when considering individual and multiple objects.

Any noise due to flickering of frames or noises added to the camera feed are effectively removed from being included in the foreground. The merging of foreground objects that might take place due to occlusion of multiple foreground objects has been avoided to a maximum extent using morphological transformations. The proposed model is a light weight model which can perform background subtraction in real time on machines with very basic processing power. Future enhancement can be shadow detection and better splitting of contours of objects that are totally occluded.

REFERENCES

[1] Arun Varghese, Sreelekha G, "Background Subtraction for Vehicle Detection," *Proceedings of Global Conference on Communication Technologies 2015 (GCCT 2015)*, pp. 380-382, 2015.

[2] Azra Nasreen, Kaushik Roy, Kunal Roy, Shobha G, "Key Frame Extraction and Foreground Modelling Using K-Means Clustering," *7th International Conference on Computational Intelligence Communication Systems and Networks (CICSyN)*, pp. 141-145, 2015.

[3] Massimo Piccardi, "Background Subtraction Techniques: A Review," *IEEE International Journal on Systems, Man and Cybernetics*, Vol. 2, (5), pp. 05-25, 2004.

[4] T. Ko, S. Soatto, D. Estrin, "Background Subtraction on Distributions," *European Conference on Computer Vision (ECCV 2008)*, pp. 222-230, October 2008.

[5] M. Himmelsbach, U. Knauer, F. Winkler, F. Zautke, K. Bienefeld, B. Meffert, "Application of an Adaptive Background Model for Monitoring Honeybees," *VIIP 2005*, 2005.
[6] Q. Ling, J. Yan, F. Li, Y. Zhang, "A Background Modelling and Foreground Segmentation Approach Based on the Feedback of Moving Objects in Traffic Surveillance Systems," *Neurocomputing*, 2014.
[7] Rahul M Patil, N R Vinay, Rohith Y, Ram Srinivas, Pratiba D, "IoT Enabled Video Surveillance System using Raspberry Pi," *2nd Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS 2017)*, December 2017.
[8] S. Chakraborty, M. Paul, M. Murshed, M. Ali, "An Efficient Video Coding Technique Using a Novel Non-parametric Background Model," *IEEE International Conference on Multimedia and Expo Workshops (ICMEW 2014)*, pp. 1-6, July 2014.
[9] X. Zhang, Y. Tian, T. Huang, W. Gao, "Low-complexity and High-efficiency Background modelling for Surveillance Video Coding," *IEEE International Conference on Visual Communication and Image Processing (VCIP 2012)*, San Jose, USA, November 2012.
[10] T. Bouwmans, "Traditional and Recent Approaches in Background modelling for Foreground Detection: An Overview," *Computer Science Review*, 2014.
[11] Gourav Takhar, Chandra Prakash, Namita Mittal, Rajesh Kumar, "Comparative Analysis of Background Subtraction Techniques and Applications," *IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2016)*, pp. 1-8, 2016.
[12] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, Deva Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, (9), pp.1627-1645, 2010.
[13] Linda Shapiro et al., editors, "Computer Vision," *Illustrated, Oxford UP, Prentice Hall*, 2001.
[14] Zezhi Chen, Tim Ellis, "A Self-Adaptive Gaussian Mixture Model," *International Journal of Elsevier Computer Vision and Image Understanding*, Vol. 122, (3), pp. 35-46, 2014.
[15] Zivkovic Z, "Improved Adaptive Gaussian Mixture Model for Background Subtraction," *Proceedings of International Conference on Pattern Recognition (ICPR)*, Moscow, pp. 28-31, 2004.
[16] Zivkovic Z, "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction and Pattern Recognition Letters," *International Journal on Pattern Recognition (IJPR)*, Vol. 27, (7), pp. 773-780, 2006.
[17] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, `http://changedetection.net`, *Proc. IEEE Workshop on Change Detection (CDW-2012) at CVPR-2012*, Providence, RI, June 2012.
[18] Faisal Bashir, Fatih Porikli, "Performance Evaluation of Object Detection and Tracking Systems," *TR2006-041, Mitsubishi Electric Research Laboratories*, 2016.
[19] Haixia Wang, Li Shi, "Foreground Model for Background Subtraction with Blind Updating," *IEEE International Conference on Signal and Image Processing*, pp. 74-78, 2016.
[20] A B Godbehere, Matsukawa A, Goldberg K, "Visual Tracking of Human Visitors Under Variable Lighting Conditions for a Responsive Audio Art Installation," *American Control Conference (ACC)*, pp. 4305-4312, June 2012.
[21] Yu Yuan, Feng D, Yuzhuo Zhong, "Fast Adaptive Variable Frame Rate Coding," *IEEE Vehicular Technology Conference*, Vol. 5, pp. 2734-2738, May 2004.
[22] Guarangnella C, Di Sciasco E, "Variable Frame Rate for Very Low Bit Rate Video Coding," *10th Mediterranean Electrotechnical Conference*, Vol. 2, pp. 503-506, 2000.
[23] Zheng Yi, Fan Liangzhong, "Moving Object Detection Based on Running Average Background and Temporal Difference," *Proceedings of International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Taiwan, pp. 270-272, 2010.
[24] Vinod M, Sravanthi T, Brahma Reddy, "An Adaptive Algorithm for Object Tracking and Counting," *International Journal of Engineering and Innovative Technology (IJEIT)*, Vol. 2, (4), pp. 560-585, 2012.