An Artificial Approach of Video Object Action Detection by Using Gaussian Mixture Model

Sneha Jain[1], Gagan Vishwakarma[2] ,Yogendra Kumar Jain[3]

Department of Computer science & Engineering,Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh (464001), INDIA

Abstract*— *The digital data processing plays an important role in today's world. This attracts many researchers to solve different issues of video object processing field. This paper works on one of significant issue of video objects detection with action recognition in videos and it proposes an artificial approach to detect action of the video object. The working methodology is consisting of the separation of foreground, background features in videos and the training of the system. Foreground and background features of the videos are separated by using gaussian mixture model. The training and action recognition of detected object is done by error back propagation neural network. The experiments are carried out on real datasets, which further calculate results of comparisons. The comparisons are taken out between experimental results and existing action detection methods. Results show that proposed work reduces the execution time and increases the pixel localization parameter.

Keywords — *Action detection, Feature extraction, Image processing, Object tracking, Neural network.*

I. INTRODUCTION

The computer vision provides different applications, out of such many applications, human action detection is highly important and attention seeker in these days. The detection system of proposed method automatically identifies the action performed by the human like running, moving, kicking, etc. So this help in monitoring the sensitive area where it automatically generates alarm for some kind of unfair activities.

The potential applications of human action detection include film and television content analysis, video index and summarization, real-time active object monitoring for video surveillance, and on-line pedestrian detection for smart vehicles.

However, human action detection remains a challenging problem. First issue in this work is appearance and body shape of the detecting object. The body shape is different for various angles of the observer.

Second issue, which affects the tracking analysis is highly varying background with illumination make it quit tough for judging the movement of the object. One more point in this issue is the moving camera which makes variable image for the same object and background. Third issue is that moving object does not repeat actions in same manner. Either the change in velocity or change in angle of movement occurs, which makes it difficult to judge.

Fourth issue include multiple objects or human being in the same scene where each perform its own activity then disocclusion and occlusion of the objects occurs which make it more difficult to judge the shape of the object with their action.

So researchers face these problems of how to extract and characterize behaviour from some video having multiple movement with multiple objects. The other motive is how to learn an efficient classifier to recognize a given behaviour in a new context. With respect to those mentioned difficulties, the main challenge is to find a set of features that characterize behaviours well and account for most of those scenarios.

Problems in visual tracking

Video object detection in general is quit challenging work because of information loss by the projection of the 3D view to 2D view. This conversion increases the noise in the frames, background illumination, full or partial occlusion, etc. Few decades before it was assumed that object movement is smooth while background is constant with no abrupt changes. But lots of improvements are done where abrupt changes are handled by leaving drifting and scenes etc. Detection of human behaviour is current issue in the working approaches and where it is needful to mention the action of detected object. So learning of these actions is highly required for different type of actions.

1. Robustness

Robustness means that even under complicated conditions, the tracking algorithms should be able to follow the interested object. The tracking difficulties may be cluttered back-ground, partial and full changing illuminations, occlusions or complex object motion.

2. Adaptively

Additional to various changes in the environment where an object is located, the object itself also undergoes with some changes. It requires a steady adaptation mechanism of the tracking system to the actual object appearance. Although it has been

studied for dozens of years, object detection and tracking remains an open research problem. A robust, accurate and high performance approach is still a challenge. The difficulty level of this problem highly depends on how the object is defined which is to be detected and tracked.

II. LITERATURE REVIEW

Rodriguez et al. [1] find that space for identifiable crowd is quite small as compared to entire space of the all possible number of crowd. This work has identified some used videos from the internet having large number of crowd. During testing, crowd patches are matched to the database in a similar fashion to that done for data-driven image de-noising and in painting. Therefore, this method requires extensive searching of similar patches in the database, while making a strong assumption that the motion of individuals in a particular query patch can be found in the database.

Fan et al. [2], propose to reduce the number of false positives by representing abandoned object alerts by relative attributes, namely statics, fore roundness and abandonment. The comparative power of these features is enumerated using a position purpose learnt on low-level temporal and spatial attributes. With the help of these attributes authors apply a linear ranking algorithm in order to sort detected action as per the relevance of the stored features.

Chatfield et al. [3], show that Convolution Neural Networks lead to significant better results than other coding methods on static image datasets. To the best of the available knowledge, this approach has not yet been evaluated for action classification or event detection in video sequences.

Wang et al. [4,] collectively used thick routes and motion borders to construct exploit descriptors, i.e., a cross between the technique of this part and those of preceding part.

Viswanath et al. [6], initiate a narrative method to notice prominent motion depends on the idea of "observability" from the productive pixels, when the frame series is signify as a linear dynamical organization. The collection of productive pixels with highest saliency is additionally used to replica the holistic dynamics of the salient area. The pixel saliency map is strengthen by two area saliency maps, which are calculated depends on the likeness of dynamics of the different spatiotemporal patches in the video with the salient region dynamics, in a global as well as a local sense. The resultant work is tested on a position of demanding series and evaluated to state-of-the-art technique to display case its better presentation on justification of its calculation efficiency and ability to notice salient motion.

Idrees et al. [11], depends completely on the information gathered from the frame of the video. This utilizes the prominent feature of the individual object in the video frames. These are very easy to trace and develop a model on that movement of the individuals from the dense crowd. It identifies the forecasted motion of the individual based on its neighbour. In order to find the motion of the individual from the crowd this work use leverages from the consecutive five frames which help in identifying the dynamic behaviour of the detecting object.

Zhong Zhou et al. [12], found algorithm, in which inherentlytemporal smoothness of actions of humans are exploited to facilitate the segmentation. Then, on the SVM framework, segmentation results of spatial and temporal extents of actions of interest are treated as latent variables and then variables are inferred simultaneously with action recognition.

Sivabalakrishnan et al. [13], used a dynamic behaviour of the individual for fuzzy based detection of the motion. In the proposed method, background is removed from the foreground for filtering objects. To gain the confidence of the moving object extracting and identification is done separately by fuzzy logic inference.

Elgammal et al. [14], has worked on the kernel density estimation, where proposed a Kernel density estimation (KDE) which is an example of non parametric technique. This proposed approach is able to solve the parametric selection property of MOG with other type of parametric properties. Although in the presence of the dynamic background scene detection of exact background is not possible by the Gaussian function. So this proposed KDE model has overcome this problem by estimating the background probabilities from current samples in the kernel density.

Yang et al. [15], executed the temporal and spatial video partition and developed a novel area based descriptor called "Motion Context" this describe both appearance and motion data of the temporal-spatial segment. In order to identify the anomaly behaviour of the object work identifies the abnormal behaviour of the event from the similar problem. This increase the robustness of the statics where model technique is used where training dataset is of small size. For the testing of the temporal-spatial segment work has trained b its best matching video. To speed up the search process, compact random projections are also adopted.

III. PRPOSED WORK

Video tracking and analysis of object recognition is a process which has few steps of methodology. Firstly the video is to be read from the dataset library and then some pre-processing steps are to be applied on the video data. These are discussed below, as under the headings of pre-processing.

1. Pre-Processing

In this step video is read in the working environment of the tool and converted into fixed size frames. So reading of frames is done by the matrixes where frames are stored. As video may be in color format means in RGB format. Then before detection, it needs to be converted into gray scale format. RGB contains three matrixes which are transformed into equivalent gray scale single matrix. One more work

is done in this step is light balancing. As most of the cameras have a mode where everything is automatically determined i.e. the white balancing and gain. It creates a noticeable amount of affect in the exposure. If something in the scene changes, the camera tries to adjust settings to give a pleasant image. Here mean of the frame is set in such a way that it becomes zero despite changes the changes which may occur in brightness and intensity.

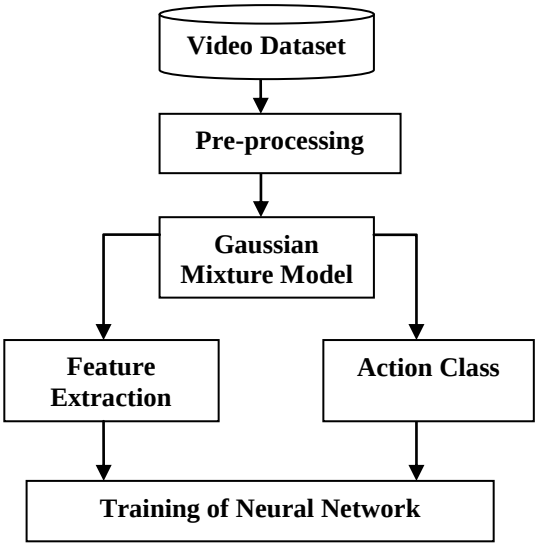

Fig.1 Block diagram of proposed model

2. Gaussian Mixture Model

The parameter of gaussian mixture model is used to separate the foreground and background features of the video.

Staufer and Grimson [5] have proposed an adaptive parametric GMM to reduce the effect of small repetitive motions like trees and buses as well as illumination variation. A pixel I at position x and time t is modelled as a mixture of K Gaussian distributions. The current pixel value follows the probability distribution given by

$$P(I_{t,x}) = \sum_{i=1}^{k} w_{(t-1,x,i)} * \eta(I_{t,x}, \mu_{(t-1,x,i)}, \sigma^2_{(t-1,x,i)})$$

where η is the Gaussian probability density function

$$\eta(I\ ,\mu\ ,\sigma^2) = \frac{1}{\sqrt{2\Pi\sigma^2}} \exp\left(-\frac{(I-\mu^2)}{2\sigma^2}\right)$$

While w, μ ,σ^2 are weight, mean value and variance of the i[th] Gaussian in the mixture at time t-1. For maintaining the Gaussian mixture model, the parameters w, μ ,σ^2 needs to be updated based on the new pixel $I_{t,x}$. A pixel is said to be matched, if $I_{t,x}$ lies within σ standard deviations of a Gaussian. In our case σ^2 lies between 1 and 5. If one of the K Gaussian parameter is matched, the matched Gaussian is updated as follows

$$\mu_{(t,x,i)} = (1-\rho)\mu_{(t-1,x,i)} + \rho(I_{t,x})$$

$$\sigma^2_{(t,x,i)} = (1-\rho)\sigma^2_{(t-1,x,i)} + \rho(I_{(t,x)}, \mu_{t,x,i})^T (I_{(t,x)}, \mu_{t,x,i})$$

where,

$$\rho = \alpha\eta(I_{t,x} \mid \mu_{t-1,x,i}, \sigma_{t-1,x,i})$$

is a learning rate that controls how fast converges to new observations.

The weight of the K Gaussian is adjusted as follows:

$$w_{t,x,i} = (1-\alpha)w_{t-1,x,i} + \alpha(M_{t,i})$$

Where $M_{t,i} = 1$ is set for the matched Gaussian and $M_{t,i}$= 0 for the others. The learning rate is used to update the weight and its value ranges between 0 and 1.

If none of the K Gaussian component matches the current pixel value, the least weighted component is replaced by a distribution with the current value as its mean, a high variance, and a low value of weight parameter is chosen. Thereafter, the weights are normalized.

The K distributions are sorted in descending order by w/σ. This ordering moves the most probable background with high weight and low variance at the top. The first B Gaussian distributions which exceed certain threshold T are retained for the background distributions. If a small value of T is chosen, the background model is uni-modal and is multi-modal, if higher value of T is chosen. If a pixel $I_{t,x}$ does not matches with any one of the background component, then the pixel is marked as foreground.

3. Feature Extraction and action class

The background is separated by applying the Gaussian mixture model on the frames of videos. In this step, pixel positions of the foreground are store in feature vector with both ‘x’, ‘y’ coordinates. This can be understood as that the pixel positions of the foreground from each frame are stored in a fixed size vectors which act like the shape of the object. As in training phase the class of the action performed by an object is known and proper class number is inserted into the output vector. So each feature vector corresponds to the class for training.

The process works in the following manner:

```
Loop 1:m // m is number of pixel in horizontal part of frame
  Loop 1:n // n is number of pixel in vertical part of frame.
        If F(m,n) is foreground
                Feature_vector(1, count)<=m
                Feature_vector(1, count+1)<=n
                Count<=count+1
                Class_vector<=C
        Endif
  EndLoop
EndLoop
```

4. Training of Error Back Propagation Neural Network (EBPNN)

When the frames of video are stored in matrixes of pixels and the features are extracted from the pixels. The next step is to detect the object and recognize the action of object. This needs a training of the system, so that the system can learn and when the experiments are being performed, it can easily detect and recognize the action of object.

Here in proposed methodology the training is done by error back propagation neural network and it has some steps which are discussed below:

- Let us assume a four layer neural network.
- Now consider i as the input layer of the network. While j is consider as the hidden layer of the network. Finally k is consider as the output layer of the network.
- If w_{ij} represents a weight of the between nodes of different consecutive layers.
- So the output of the neural network is depend on the below equation:

$$Yj = \frac{1}{1+e^{-X_j}}$$

where, $X_j = \sum x_i \, . \, w_{ij} - \theta_j$, $1 \leq i \leq n$, n is the number of inputs to node j, and θ_j is threshold for node j

- The error of output neuron k after the activation of the network on the n-th training example (x(n), d(n)) is:

$$e_k(n) = d_k(n) - y_k(n)$$

- The network error is the sum of the squared errors of the output neurons:

$$E(n) = \sum e_k^2(n)$$

- The total mean squared error is the average of the network errors of the training examples.

$$E_{avg} = \frac{1}{N} \sum_{n=1}^{N} E(n)$$

- The Back propagation weight update rule is based on the gradient descent method. It takes a step in the direction yielding the maximum decrease of the network error E. This direction is the opposite of the gradient of E.
- Iteration of the Back propagation algorithm is usually terminated when the sum of squares of errors of the output values for all training data in an epoch is less than some threshold such as 0.01

$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad \Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

IV. EXPERIMENTS AND RESULTS

In order to implement above algorithm for object action detection system MATLAB 2012A is used.

All algorithms and utility measures are implemented using the MATLAB tool. The tests are performed on a 2.27 GHz Intel Core i3 machine, equipped with 4GB of RAM, and running under Windows 7 Professional.

Neural Network Toolbox includes command-line functions and applications for creating, training, and simulating neural networks. This makes it easy to develop neural networks for tasks such as data-fitting, pattern recognition, and clustering.

Dataset

To perform the experiments the real dataset is used. Here video files are of different actions with same frame dimension of 120x160. Each video is of .avi format.

Evaluation Parameter

When experiments are performed, there is a need to characterize some parameters, so that on the basis of parameters the working efficiency of methodology can be analysed and these parameters are known as evaluation parameters.

- **Execution time**

This is the time taken by the algorithm to detect the action in the video or it can also be said in terms of the total time taken by the system, which includes the video reading time and execution time completely.

- **Action Localization**

This parameter of evaluation is defined as the capacity of system to locate pixels. When a video is read and processed for object detection, at the time of execution the pixels are plotted according to the actions of object.

Table I COMPARES EXECUTION TIME

Actions	Execution time in seconds	
	Proposed Work	Previous Work [12]
Walking	6.96	22.34
Clapping	5.95	25.59
Waving	7.02	24.81
Running	6.34	30.15

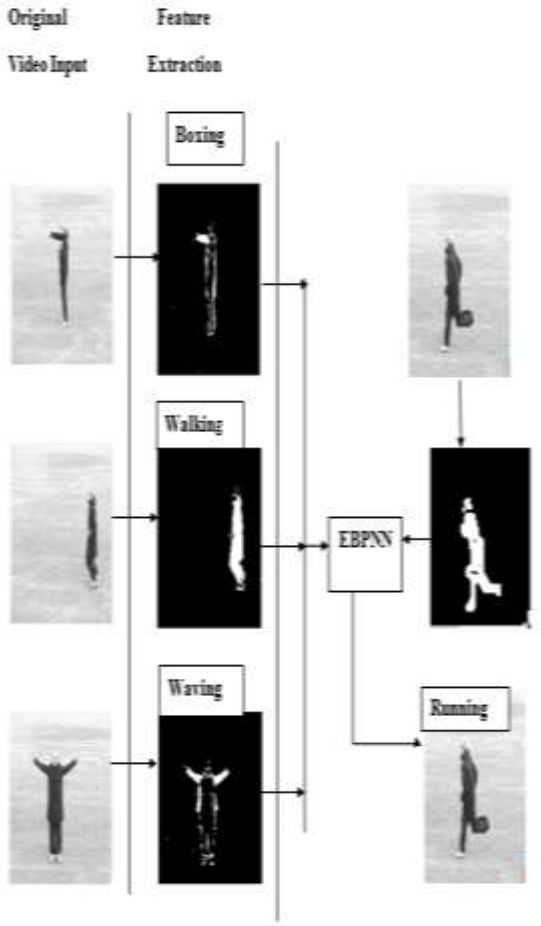

Fig. 2 Working methodology of algorithm

Table I shows that proposed work has highly decrease the execution time. The processing time is lesser in proposed methodology so the total time of execution becomes less as compared to previous strategy.

TABLE II
COMPARISION OF PREVIOUS AND PROPOSED WORK ON ACTION LOCALIZATION OF PIXELS

Actions	Action localization pixels	
	Previous Work [12]	Proposed work
Jogging	1586.9	844.7
Boxing	1280.3	387.6
Waving	841	457
Running	1956.2	817.9

The description in Table II puts an efficient analysis of action localization parameter. It shows that the proposed work has highly decrease the action localization pixels. The decrement concludes that the object is localized and tracked in accurate way.

TABLE III
COMPARISON OF PROPOSED AND PREVIOUS WORK ON ACTION

Actions	Human Action	
	Previous Work [12]	Proposed work
Walking	Moving	Walking
Boxing	Standing	Boxing
Waving	Standing	Waving

Table III shows that proposed work has accurately detect the actions video. Here due to use of neural network for testing detection of action is quit feasible.

The comparative results of previous work and proposed work is shown by plotting the execution time parameter and action localization parameter. In Fig. 3, An experiment of execution is shown, in which proposed work plots pixels of action localization more accurate than previous work.

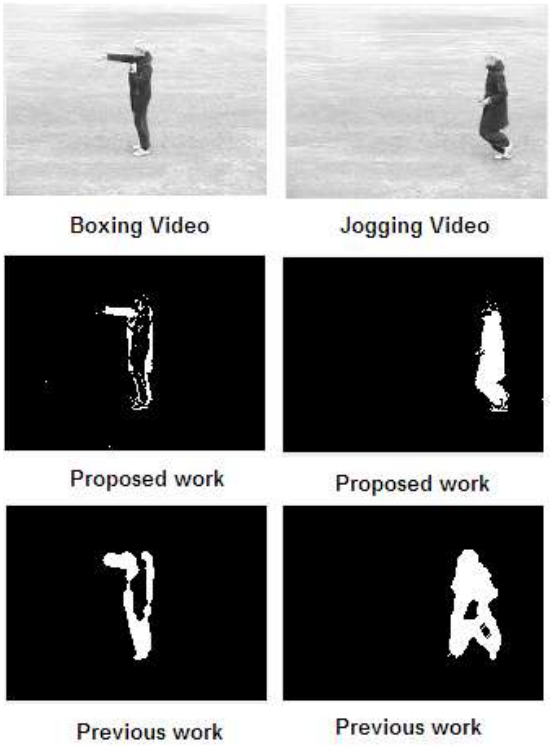

Fig. 3 Results obtains after testing videos.

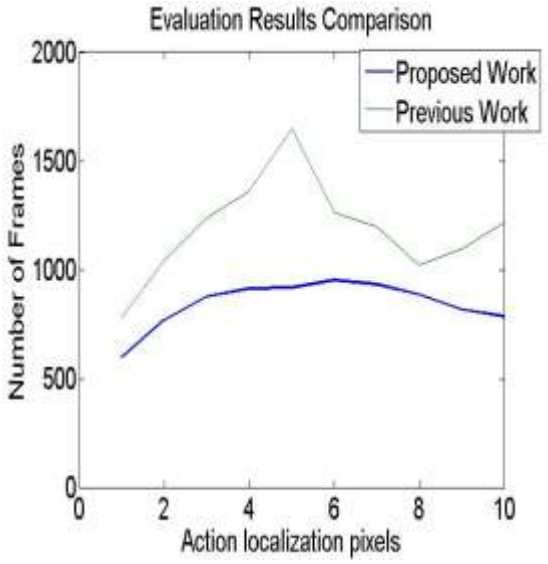

Fig. 4 action localization pixel values on each frame

Graph in Fig. 4 shows the evaluation results in such a way that the proposed work identifies the action in each frame comparatively in a better way.

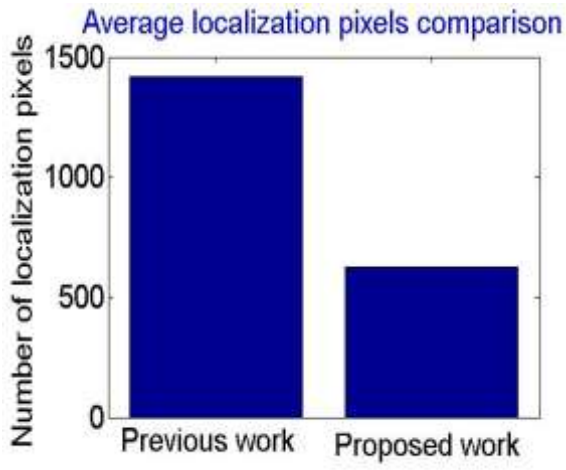

Fig. 5 Average pixel localization comparison graph.

Fig. 5 shows that proposed work has decrease the localization pixels for object detection as compare to previous work. Graph shows that average localization number of pixels for all kind of video in proposed work is less while testing.

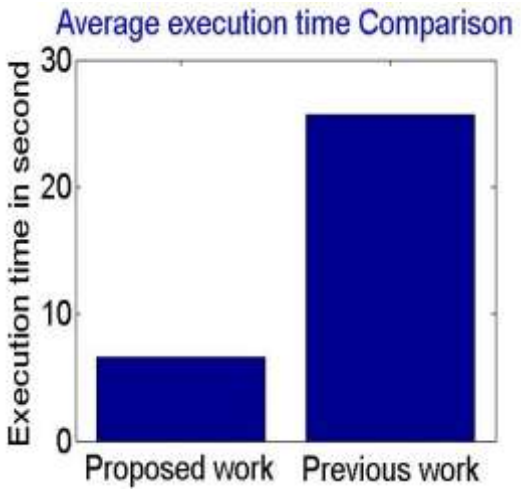

Fig. 6 Average execution time comparison graph

Fig. 6 shows that the average execution time is less in case of proposed work as compared to previous work.

While analysing these evaluation parameters it is found as that the previous work uses SVM for training and the SVM technique uses two region classification. So this becomes a limitation in case of video object action detection system because all action category need to be defined in those regions only.

In the proposed methodology the GMM technique is used, which has a distribution function so the features are extracted with a wider range of pixel locations and the use of neural network for testing of actions detection is quit feasible.

V. CONCLUSION

The video object detection plays an important role in surveillance system. Contributing in this field of video object action detection is done in this paper. The key idea is to separate the foreground and background pixels in the frame by using Gaussian mixture model. Here object pixel patterns are taken as the training feature input for the neural network. Trained neural network by this Gaussian mixture makes efficient video object action detection system. Values obtained from different evaluation parameters show that the proposed work reduces execution time by 3.91 seconds. Results show that multiple actions are detected from the same trained neural network. Proposed work considers the dynamics of the video sequence and it does not consider spatial information. Future work will involve integrating spatial information into this framework.

VI. REFERENCES

[1]. Rodriguez, M., Sivi. C., J., Laptev, I., and Audibert, J.Y., "Data-Driven Crowd Analysis In Videos", 13th International Conference On Computer Vision, Barcelona, ICCV 2011, Pages1235-1242, 2011.

[2]. Fan, Q., Gabbur, P., and Pankanti, S., "Relative Attributes For Largescale Abandoned Object Detection", 14th International Conference On Computer Vision, Australia, Pages 2736–2743, 2013.

[3]. Chatfield, K., Simonyan, K., Vedaldi, A., And Zisserman, A., "Return Of The Devil In The Details: Delving Deep Into Convolutional Nets" Technical Report, University Of Oxford. Archived In Arxiv Pages 1405-3531, 2014.
[4]. Wang, H., Kl Aser, A., Schmid, C., And Liu, C.L., "Dense Trajectories And Motion Boundary Descriptors For Action Recognition", International Journal Of Computer Vision, Vol. 103(1), Pages 60-79, 2013.
[5]. C. Stauffer And W. E. L. Grimson, "Adaptive Background Mixture Models For Real-Time Tracking In Computer Vision And Pattern Recognition", IEEE Computer Society, Pages 2246–2252, 1999.
[6]. Viswanath Gopalakrishnan, Deepu Rajan, And Yiqun Hu, "A Linear Dynamical System Framework For Salient Motion Detection", IEEE Transactions On Circuits And Systems For Video Technology, Vol. 22, NO. 5, Page 683, MAY 2012.
[7]. W.T. Lee And H. T. Chen, "Histogram-Based Interest Point Detectors", IEEE Conference On Computer Vision And Pattern Recognition, Pages 1590-1596, 2009.
[8]. Rupesh Kumar Rout, "A Survey On Object Detection and Tracking Algorithms", Department Of Computer Science And Engineering National Institute Of Technology Rourkela Rourkela 769 008, India.
[9]. Jiuyuehao, Chao Li, Zuwhan Kim, and Zhang Xiong, "Spatio-Temporal Traffic Scene Modeling For Object Motion Detection", IEEE Intelligent Transportation Systems, 2012.
[10]. Liu Gangl, Ningshangkun, You Yugan, Wen Guanglei And Zhengsiguo, "An Improved Moving Objects Detection Algorithm", IEEE International Conference On Wavelet Analysis And Pattern Recognition, Pages 96-102, 14-17 July, 2013.
[11]. Idrees, H., Warner, N., And Shah, M., "Tracking In Dense Crowds Using Prominence And Neighborhood Motion Concurrence", Image And Vision Computing, Vol. 32(1), Pages 14–26, 2014.
[12]. Zhong Zhou, Member, Feng Shi, And Wei Wu., "Learning spatial And Temporal Extents Of human Actions For Action Detection", DOI 10.1109/TMM.2015.2404779, IEEE Transactions On Multimedia, 2015.
[13]. Sivabalakrishnan, M., Manjula D., "Adaptive Background Subtraction In Dynamic Environments Using Fuzzy Logic", International Journal On Computer Science And Engineering, Vol. 2, Pages 270–273, 2010.
[14]. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L., "Background And Foreground Modeling Using Nonparametric Kernel Density Estimation For Visual Surveillance", IEEE 90, Pages 1151 – 1163, 2002.
[15]. Yang Cong, Junsong Yuan, and Yandong Tang, "Video Anomaly Search in Crowded Scenes via Spatio-Temporal Motion Context", IEEE Transactions On Information Forensics And Security, Vol. 8, No. 10, October 2013.