INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 13(2), 162-174, April 2018.

Moving Object Detection and Tracking using Genetic Algorithm Enabled Extreme Learning Machine

G. Jemilda, S. Baulkani

G. Jemilda*
Jayaraj Annapackiam CSI College of Engineering, Nazareth, India.
*Corresponding author: jemildajeba@yahoo.com

S. Baulkani
Government College of Engineering, Tirunelveli, India.
ramabaulkani@yahoo.co.in

Abstract: In this proposed work, the moving object is localized using curvelet transform, soft thresholding and frame differencing. The feature extraction techniques are applied on to the localized object and the texture, color and shape information of objects are considered. To extract the shape information, Speeded Up Robust Features (SURF) is used. To extract the texture features, the Enhanced Local Vector Pattern (ELVP) and to extract color features, Histogram of Gradient (HOG) are used and then reduced feature set obtained using genetic algorithm are fused to form a single feature vector and given into the Extreme Learning Machine (ELM) to classify the objects. The performance of the proposed work is compared with Naive Bayes, Support Vector Machine, Feed Forward Neural Network and Probabilistic Neural Network and inferred that the proposed method performs better.
Keywords: curvelet transform, speeded up robust features, enhanced local vector pattern, histogram of gradient, extreme learning machine, genetic algorithm.

1 Introduction

Tracking an object (or multiple objects) on an image plays a vital role in the field of computer vision. Numerous applications of object tracking in computer vision are video compression, video surveillance, visual interface, man-and-computer management, medicine, augmented reality and robotics. Quick movements of objects change shapes, scenes, unstructured structures of objects and cameras have trouble tracking objects. Tracking is normally performed in the context of a program that requires the location and/or shape of the object in each frame. One of the most advanced algorithms for feature selection is the genetic algorithm. This is a stochastic method for function optimization based on the mechanics of natural genetics and biological evolution.

Faheem and Gungor [5] proposed a novel dynamic clustering based energy efficient and QoS aware routing protocol (called EQRP) which is made by the real behavior of the bird mating optimization (BMO) for smart grid applications. Fadel et al. [3] introduced Cardio radio sensor to serve as a reliable, robust and efficient communication address. In this paper, honey bee mating and cooperative channel assignment algorithms have been proposed. This significantly decreases the probability of packet loss, energy consumption, improving the life time of CRSNs in smart grids. Faheem et al. [4] discussed a novel nature-inspired evolutionary link quality-aware queue-based spectral clustering routing protocol for UASN-based underwater applications. To overcome the challenges faced by the commonly used UWSN performance indicator and to overcome the inefficiencies existing clustering based routing protocol, a novel QoS aware evolutionary cluster based protocol mentioned by Faheem et al. [6]. The main advantage is that it is robust and efficient. It is used for optimization, searching, feature extraction, segmentation and classification. This genetic algorithm reduces the processing time and increases the accuracy.

The main contributions of this paper can be described as follows:

- The use of color information (instead of grey levels only), during the entire detection process, which improves sensitivity;

- The use of background subtraction function is a good approximation of a mode function and assures high responsiveness to background changes and good accuracy;

- The inclusion of knowledge-based feedback from the soft thresholding segmentation module abates false positives and prevents from deadlock situations in background update;

- The system is conceived to be auto-adaptive, i.e., to work with the best possible performances on all possible scenarios. Results show that accurate results are aligned with other state-of-the-art approaches, but they are more stable in all corner-case situations and higher frame-per-second rates.

- It also shows very limited memory (of all types) requirements, as it is usually demanded for embedded applications.

The rest of this paper is organized as follows: Section 2 discusses the related work. Section 3 explains the proposed method including its design idea and practical implementation approach. Section 4 presents simulation parameters. Section 5 provides experimental results where the effectiveness of the proposed work is compared to the existing methods. Finally, Section 6 concludes this paper by summarizing our results, significance and future possibilities of the work.

2 Related work

A new tool developed by Arnab Roy et al. [12] to speed up the dramatic change of detection of low-end laptop webcam stuff with specific hardware for processing high-speed images. The first algorithm is fast and ensures the second algorithm edge of the object found more clearly at the expense of slow image processing. This method is used to reduce the average delay in motion of the objects up to 45.5% and the memory usage by about 14%, while keeping the same accuracy. Hsu-Jung Cheng et al. [2] proposed a framework for detecting vehicles in aeronautical surveillance using the Dynamic Bayesian Network and found that this method has flexibility and good generalization capabilities. Felix M. Philip et al. [11] proposed visibility model based on the second derivative to predict the objects, spatial tracking model and tangential weighted function to track multiple objects. But tracking multiple objects in low resolution videos is not possible.

3 Methodology

This section briefly explains about the methods used in this paper for tracking the objects and it was shown in the Figure 1.

3.1 Input video

The input video is chosen as it consists of the objects to track. The objects in the video must be moving, since this project is to capture the moving objects in the video.

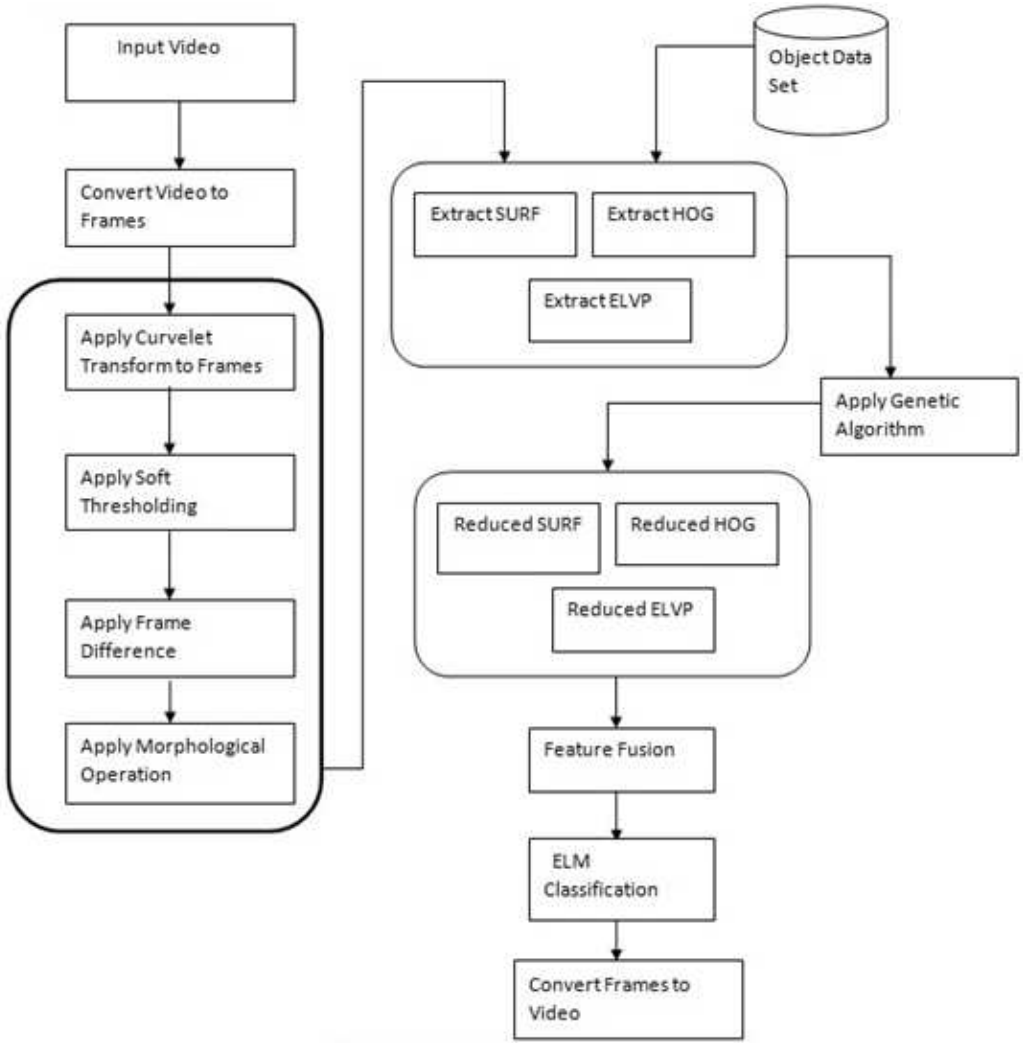

Figure 1: Architecture diagram

3.2 Divide video into image frames

Video technology is widely used for electronically capturing, recording, processing, storing, transmitting and analyzing a sequence of still scenes showing images in motion. The number of still pictures of videos per unit of time ranges from six or eight frames per second (frame/s) to 120 frames per second or more. These frames are saved and processed consequently. In this context, the video is divided into number of frames for example, the video of up to 10 seconds is break down into 80 to 85 frames.

3.3 Object localization

The moving object is localized from each and every frame by using background subtraction algorithm. The traditional background subtraction algorithm is modified and a new concept is implemented. The curvelet transform is applied on each and every frame. Then soft thresholding is applied to remove noise. Next, frame difference process is applied to segment the moving object and the background. Finally, mathematical morphology concept is applied to fill the gap and edges in the detected objects.

Curvelet transform

Curvlet is not a coordinate technique for presenting multiple objects. It differs from the other wave in the direction of transition to the extent of oriented translation. The use of this transform is optimally sparse representation of objects with edges and optimal image reconstruction in ill-posed problems [9].

Soft thresholding

Soft thresholding is a popular tool in computer vision and machine learning. It is used to remove noise from an image [1].

Frame difference approach

Frame differencing is a technique used to find the difference between two video frames. It is used to segment the moving object from the background. If the pixels have changed, apparently there was something changed in the image. The algorithm of frame difference is relatively simple. It is described in very detailed manner as follows:
1. Convert the incoming frame to grayscale.
2. Subtract the current frame from the background model, which is the previous frame.
3. For each pixel, if the difference between the current frame and background is greater than a threshold then the pixel is considered as part of the foreground. Otherwise it is considered as the background.
4. The foreground pixel is denoted as white color and the background pixel is denoted as the black color. Finally the output image is produced.

Mathematical morphology

Mathematical morphology is a powerful tool for extracting structural characteristics in an image and is useful for characterizing shape information.The obtained segmented object may include a number of disconnected edges due to non-ideal segmentation of moving object edges. Therefore, some morphological operation is needed to generate connected edges. Here, a binary closing morphological operation is used and it is determined by a structuring element. The effect of the operator is to preserve background regions that have a similar shape to this structuring element, or that can completely contain the structuring element, while eliminating all other regions of background pixels [8].

3.4 Object identification

After locating the moving object, the next step is to identify the object. To do this first the feature extraction techniques are applied on to the localized object. In this paper the texture, color and shape information of objects are considered.

Shape feature extraction using SURF

Speeded Up Robust Features (SURF) is a local feature detector and descriptor that can be used for tasks such as object recognition, registration, classification and 3D reconstruction. It is partly inspired by the scale-invariant feature transform (SIFT) descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT. SURF descriptors can be used to locate and

recognize objects, people or faces, to make 3D scenes, to track objects and to extract points of interest [15].

SURF uses the determinant of Hessian for selecting the scale, as it is done by Lindeberg. Given a point p=(x, y) in an image I, the Hessian matrix H (p, σ) at point p and scale σ, is defined as follows:

$$H\left(p,\sigma\right)=\begin{pmatrix} L_{xx}\left(p,\sigma\right)L_{xy}(p,\sigma) \\ L_{x,y}\left(p,\sigma\right)L_{yy}\left(p\ ,\sigma\right)\end{pmatrix} \tag{1}$$

where $L_{xx}\left(p,\sigma\right)$ etc. are the second-order derivatives of the grayscale image. The box filter of size 9×9 is an approximation of a Gaussian with σ=1.2 and represents the lowest level (highest spatial resolution) for blob-response maps. The pseudo code for SURF is shown in Algorithm 1.

Algorithm 1: Pseudo Code of SURF algorithm
Input : Input Image Output : SURF Key Points
1. Begin 2. For all pixel i=1 : P 3. do 4. Compute the gradient. 5. Compute gradient magnitude and orientation 6. Find the scalespace response using DoH filters with different σ. 7. Find local maxima LM with different scales and octaves. 8. if LM==Highest Maxima value 9. Select as SURF Key Points 10. end if 11. end 12. End

Colour feature extraction using HOG

The histogram of oriented gradients (HOG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. It is used to count the occurrences of gradient orientation in localized portions of an image [10]. The pseudo code for HOG is shown in Algorithm 2.

Algorithm 2: Pseudo Code of HOG algorithm
Input : Input Image Output : HOG Feature
1. Begin 2. For all block i=1 : B 3. do 4. Compute gradient values in horizontal directions. 5. Compute gradient values in vertical directions. 6. Create histograms of gradient image 7. end 8. Concatenate normalized histograms from all of the block regions. 9. End

Texture feature extraction using ELVP

Enhanced Local Vector Pattern (ELVP) is a novel vector representation developed to represent the 1D direction and structure information of local texture and the adjacent pixels with

diverse distances from different directions. Based on the vector representation, the LVP descriptor is proposed to provide various 2D spatial structures of micro patterns with various pair wise directions of vector of the referenced pixel and its neighborhoods.

The proposed local pattern descriptor encodes the LVPs by using the four pairwise directions of vector of the referenced pixel and its neighborhoods. Especially, each pairwise direction of vector of the referenced pixel generates the transform ratio which is used to design the weight vector of dynamic linear decision function for encoding the distinct 8-bit binary pattern of each LVP. Since the binary code can be considered as a two-class case by using dynamic linear decision function to calculate the Comparative Space Transform (CST) values of the neighborhoods for encoding a bit string via the sign function [7]. The pseudo code for ELVP is shown in Algorithm 3.

Algorithm 3: Pseudo Code of ELVP algorithm
Input : Input Image Output : ELVP Feature
1. Begin 2. For all pixel i=1 : P 3. do 4. Take the one pixel G_c 5. Take the pixel N in direction of angle β 6. Calculate the Difference D between G_c and N $V_{\beta,D}(G_c) = (I(G_\beta, D) - I(G_c))$ 7. Calculate the binary pattern $LVP_{p,R}(G_c) = \{LVP_{p,R,\beta}(G_c) \mid \beta = 0^o, 45^0, 90^0, 135^0\}$ 8. Calculate the binary description and store it into the bin value. 9. Store Descript ion in the bin array 10. End 11. End

3.5 Feature selection

The genetic algorithm (GA) is used for finding the best features from the SURF, HOG and ELVP. A better solution can be evolved using a population of strings (called chromosomes or the genotype of the genome) and by encoding candidate solutions (called phenotypes) to an optimization problem. Solutions are represented in binary as strings of 0s and 1s. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated. Multiple individuals are selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. The algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached. GA proceeds to initialize a population of solutions randomly, and then improve it through repetitive application of mutation, crossover, and inversion and selection operators [13].

Apply the genetic algorithm for each feature extraction method such as SURF, HOG and ELVP. After applying the genetic algorithm the reduced SURF, reduced HOG and reduced ELVP is produced.

3.6 Feature fusion

The next step is to combine the reduced SURF, reduced HOG and reduced ELVP into single feature vector. To do this, concatenate all features into a single array. Finally the single feature vector is produced.

3.7 Object classification

A new learning algorithm called extreme learning machine (ELM) is used to classify objects for single-hidden layer feed forward neural networks (SLFNs) which randomly chooses hidden nodes. This algorithm provides good performance at fast learning speed. The ELM is an emerging learning technique provides efficient solutions to generalized feed-forward networks including (both single and multi-hidden-layer) neural networks, radial basis function (RBF) networks, and kernel learning. ELMs have classification capability. ELMs have significant features like fast learning speed, ease of implementation, and minimal human intervention. They thus have strong potential as a viable alternative technique for large-scale computing and machine learning [14].

3.8 Generating video from frames

To generate a video from a set of sequences or set of frames, it starts with the number zero. This will work as long as the sequence is unbroken. The frame rate of the resulting video is 3 frames per second so that each still can be seen for a short period of time. The rescale of the picture to the desired resolution can be done to manage the size of the resulting video.

3.9 Output video

The output video is generated from the filtered and classified frames. These frames are combined to form the output video.

4 Simulation

The Simulation results on five different sequences recorded with three different platforms is presented in Table 1. In all cases, the sensor was a pair of forward-looking AVT Marlin F033C cameras, which deliver synchronised video streams of resolution 640 x 480 pixels at 13-14 frames per second. Bahnhofstrasse (999 frames) and Linthescher (1208 frames) have been recorded with a child stroller (baseline $\approx$ 0.4 m, sensor height $\approx$ 1 m, aperture angle $\approx 65^\circ$) in busy pedestrian zones, with people and street furniture frequently obstructing large portions of the field of view. Loewenplatz (800 frames), Bellevue (1500 frames) and City (3000 frames) have been recorded from a car (baseline $\approx$ 1 m, sensor height $\approx$ 1.3 m, aperture angle $\approx 50^\circ$) driving on inner-city streets among other cars, trucks and trams. Pedestrians appear mostly on sidewalks and crossings, and are observed only for short time spans. Lighting and contrast are realistic, with most sequences recorded on cloudy days.

5 Experimental images

The datasets used in this work are taken from https://motchallenge.net/vis/MOT17-13-SDP and https://motchallenge.net/vis/ETH-Crossing. Only the frames of the datasets for various movements of the moving object 1 and 2 are shown in the Figure 2 and Figure 3.

Table 1: Simulation results

Simulation Parameter	Value
Population	12
No of Generation	10
Video Resolution	640 x 480
Background updation Frame	10

Figure 2: Various movements of the moving object 1

5.1 Experimental results

To evaluate the performance of the proposed system, this paper used the dataset with various objects like moving objects, lemming, occlusion, multiple objects, and indoor objects. To examine the effectiveness of the proposed system, it is compared with various performance metrics. The various performance metrics used to compare the effectiveness are average center error, overlap rate, Detection Accuracy, Sensitivity, Specificity, F-Measure, Detection Error, Execution Time, Precision Rate and Recall Rate.

Overlap rate

It shows the overlapping of each frame by frame.

$$OverlapRate = \frac{area\ (R_T \cap R_G)}{area\ (R_T \cup R_G)} \tag{2}$$

Where R_T - Tracking result of each frame
R_G - Corresponding ground truth

Average center error rate

It shows the error rate for each frame

$$AverageCenterErrorRate = area\,(R_T \cup R_G) - area\ (R_T \cap R_G) \tag{3}$$

Where R_T - Tracking result of each frame
R_G - Corresponding ground truth

Figure 3: Various movements of the moving object 2

Precision rate

The precision is the fraction of retrieved instances that are relevant to the find.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Where TP - True Positive (Equivalent with Hit)
FP - False Positive (Equivalent with False Alarm)

Detection accuracy

It is one of the parameter which is used to analyse the performance of the proposed method.

$$\text{Detection Accuracy} = \frac{\text{No. of Correctly Classified Objects}}{\text{Total No. of Objects}} \tag{5}$$

F-measure

F-measure is the ratio of product of precision and recall to the sum of precision and recall. The f-measure can be calculated as

$$\text{F}_\text{m} = (1+a) * \frac{\text{Precision} * \text{Recall}}{a * \text{Precision} + \text{Recall}} \tag{6}$$

Where a - real value

Sensitivity

Sensitivity also called the true positive rate or the recall rate measures the proportion of actual positives. It is calculated using

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{7}$$

Where TP – True Positive (equivalent with hit)
FN - False Negative (equivalent with miss)

Specificity

Specificity measures the proportion of negatives which are correctly identified. It is calculated using

$$\text{Specificity} = \frac{\text{TN}}{(\text{FP} + \text{TN})} \tag{8}$$

Where TN – True Negative (equivalent with correct rejection)
FP – False Positive (equivalent with false alarm)

Detection error

It is one of the parameter which is used to analyse the performance of the proposed method. It is calculated using the below said formula.

$$\text{Detection Error} = \frac{\text{No. of Wrongly Classified Objects}}{\text{Total No. of Objects}} \tag{9}$$

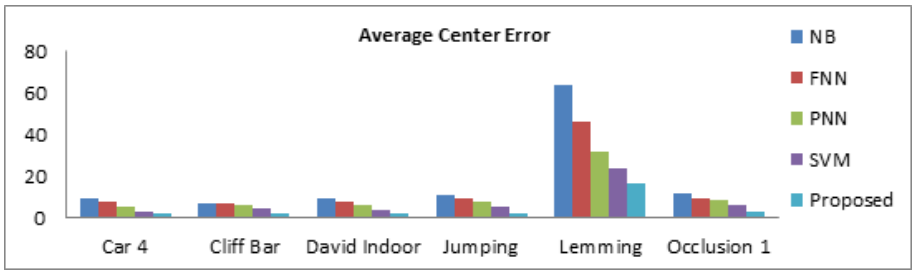

Figure 4: Average center error for various methods with different datasets

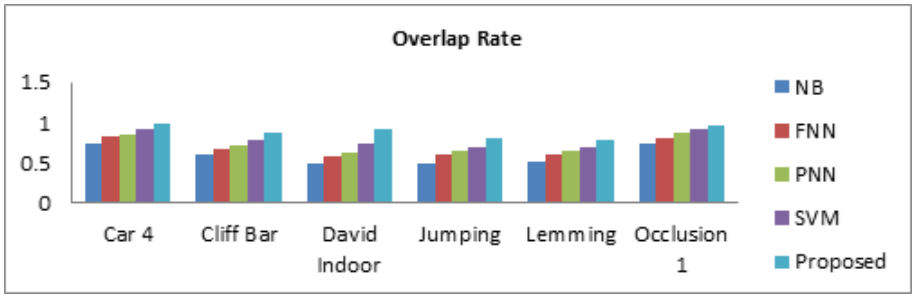

Figure 5: Overlap rate for various methods with different datasets

Execution time

It is one of the parameter which is used to analyse the performance of the proposed method. It is calculated using the below said formula.

$$\text{Execution Time} = \text{Ending Time} - \text{Starting Time} \tag{10}$$

To analyse the performance of the proposed system, it is compared with the above mentioned performance metrics. The performance comparison of the average center error value of the proposed method and other four existing approaches such as NB, FNN, PNN and SVM are plotted in the graphs as below.

In the Figure 4, the average center error value of the five methods including the proposed method is compared. The average center error value of the proposed method is lower than the other four existing approaches. Because of the low average center error value, the proposed method is better than the other four existing approaches.

The performance comparison of the overlap rate of the proposed method and other four existing approaches such as NB, FNN, PNN and SVM is compared and shown in the Figure 5. The overlap rate value of the proposed method is higher than the other methods. Because of the high overlap rate value, the proposed method is better than the other four existing approaches.

The performance precision rate of the proposed method and other four existing approaches such as NB, FNN, PNN and SVM is plotted in the Figure 6. The precision rate value of the proposed method is higher than the other four existing approaches. Because of the high precision rate value, the proposed method is better than the other four existing approaches.

The detection accuracy value of the proposed method is higher than the other four existing approaches as shown in the above Figure 7. Because of the higher detection accuracy value, the proposed method is better than the other four existing approaches.

The sensitivity value of the proposed method is higher than the other four existing approaches as shown in the above Figure 8. Because of the high sensitivity value, the proposed method is better than the other four existing approaches.

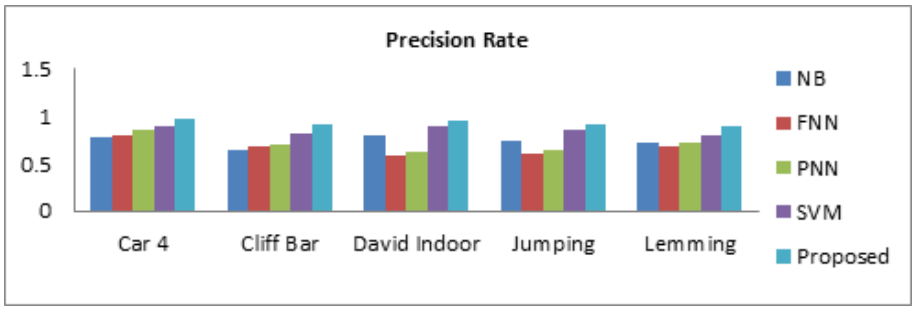

Figure 6: Precision rate for various methods with different datasets

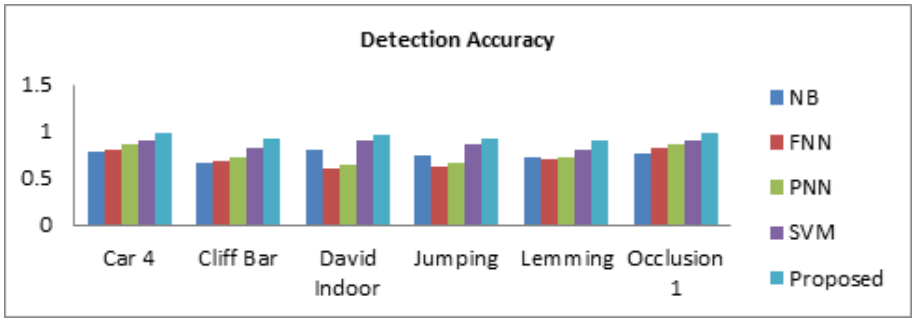

Figure 7: Detection accuracy for various methods with different datasets

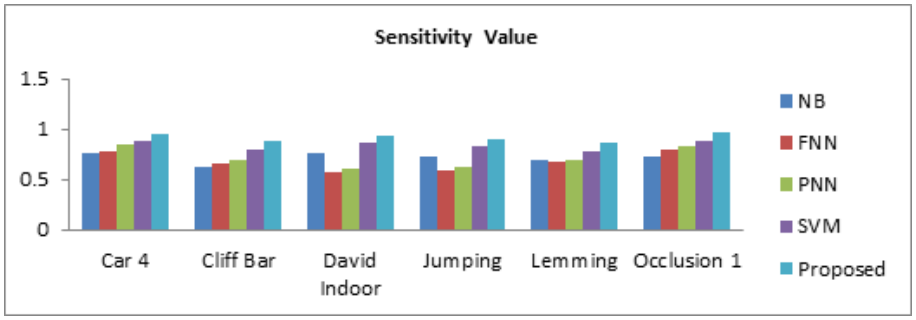

Figure 8: Sensitivity value for various methods with different datasets

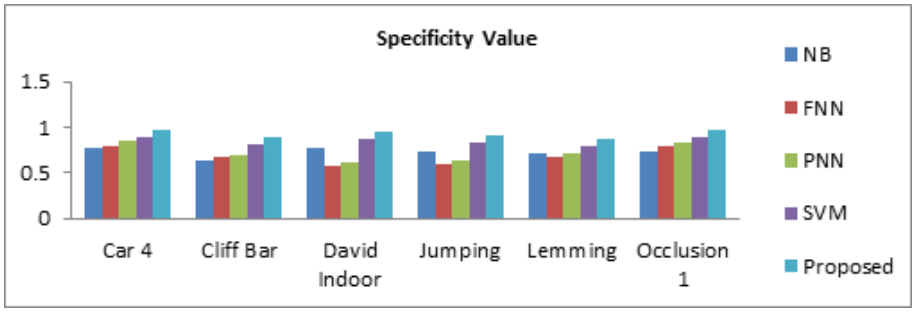

Figure 9: Sensitivity value for various methods with different datasets

The Figure 9 shows the specificity value of the proposed method is higher than the other four existing approaches. Because of the high specificity value, the proposed method is better than the other four existing approaches.

6 Conclusion

In this paper, the moving objects in the video are tracked and it is highlighted to show that the object is moving in the video. For the first the video is divided into frames. First the input given is a video, which is divided into frames. Then the moving object is localized from each and every frame by using background subtraction algorithm. After locating the moving object the next step is to identify the object. To do this first the feature extraction techniques are applied on to the localized object. After extracting all the features these feature are reduced using the genetic algorithm and then these reduced feature set are fused to form a single feature vector. Finally these feature vectors are given into the Extreme Learning Machine (ELM) to classify the objects. The classified frames are then combined to form a video. In these output video, the moving object is highlighted. To know the performance of the proposed method, we compared the results with the various methods. From the experiment results, our proposed methods shows better results for identifying the moving object in the video file and are very efficient and reliable.

As a future work, an algorithm that detects the tracking failure and recovers the tracking process when the target tracking fails due to long duration of heavy occlusion is planned.

Bibliography

[1] Biswas, M.; Om H. (2012); A new soft thresholding Image Denoising method, *Science Direct*, 6:10-15,2012.

[2] Cheng, H.-Y.; Weng, C.-C.; Chen Y.-Y.(2012); Vehicle Detection in Aerial Surveillance Using Dynamic Bayesian Networks, *IEEE Transactions on Image Processing*, 21(4): 2152-2159, 2012.

[3] Fadel, E.; Faheem, M.; Gungor, V.; Nassef, L.; Akkari, N.; Malik, M. (2017); Spectrum-Aware Bio-Inspired Routing in Cognitive Radio Sensor Networks for Smart Grid Applications *Computer Communications*, 106-120, 2017.

[4] Faheem, M.; Tuna, G.; Gungor, V.C. (2016); LRP: Link quality- aware queue- based spectral clustering routing protocol for underwater acoustic sensor networks, *International Journal of Communication Systems*, 2016.

[5] Faheem, M.; Gungor V.C. (2017); Energy Efficient and QoS-aware Routing Protocol for Wireless Sensor Network-based Smart Grid Applications in the Context of Industry 4.0, *Applied Soft Computing*, 1-13, 2017.

[6] Faheem, M.; Tuna, G.; Gungor V.C. (2017); QERP: Quality-of-Service (QoS) Aware Evolutionary Routing Protocol for Underwater Wireless Sensor Networks, *IEEE Systems Journal*, 2017.

[7] Fan, K.-K.; Hung, T.-Y. (2014); A Novel Local Pattern Descriptor-Local Vector Pattern in High-Order Derivative Space for Face Recognition, *IEEE Transactions on Image Processing*, 23(7), 2877 - 2891, 2014.

[8] Kimori, Y. (2013); Morphological image processing for quantitative shape analysis of biomedical structures: effective contrast enhancement, *Journal of Synchrotron Radiation*, 1(20), 848-853, 2013.

[9] Kourav, A.; Singh P. (2013); Review on curvelet transform and its applications, *Asian Journal of Electrical Sciences*, 2(1): 9-13, 2013.

[10] Li, Y.; Su G. (2015); Simplified histograms of oriented gradient features extraction algorithm for the hardware implementation, *International Conference on Computers, Communications and Systems (ICCCS)*, 192 -195, 2015.

[11] Philip, F.M.; Mukesh R.(2016); Hybrid tracking model for multiple object videos using second derivative based visibility model and tangential weighted spatial tracking model, *International Journal of Computational Intelligence Systems*, 9(5): 888-899, 2016.

[12] Roy, A.; Shinde,S.; Kang, K.-D. (2012); An Approach for Efficient Real Time Moving Object Detection, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 5(3), 2012.

[13] Shingade, A.; Ghotkar A.(2014); Survey of Object Tracking and Feature Extraction Using Genetic Algorithm, *International Journal of Computer Science and Technology*, 5(1), 2014.

[14] Wang, Y.; Cao, F.; Yuan, Y. (2014); A Study on Effectiveness of Extreme Learning Machine, *arXiv:1409.3924v1 [cs.NE]*, 13, 2014.

[15] Zohrevand, A.; Ahmadyfard, A.; Pouyan, A.; Imani, Z. (2014); A SIFT based object recognition using contextual information, *Iranian Conference on Intelligent Systems (ICIS)*, 1-4, 2014.