Implementation of Modified Mask RCNN

Date Archana, Shah Sanjeevani

Abstract: ***Detecting camouflage moving object from the video sequence is the big challenge in computer vision. To detect moving object from dynamic background is also very difficult as the background is also detected as moving object. Mask RCNN is a deep neural network which solves the problem of separation of instances of same object in machine learning or computer vision. Thus, it separates different objects in video. It is the extension of faster RCNN in which an extra branch is added to create an object mask simultaneously along with bounding box and classifier. After giving input, Mask RCNN gives the rectangle around the object, class to which object belong and object mask. This article introduces Mask RCNN algorithm along with some modifications for target detection from dynamic background and also for camouflage handling. After target object detection, contrast limited adaptive histogram equalization is applied. Morphological operations are used to improve results. For both challenges quantitative and qualitative measures were obtained and compared with the existing algorithms. Our method efficiently detects the moving object from input sequence and gives best results in both situations.***

Keywords: **Camouflage, object, dynamic, moving, video surveillance.**

I. INTRODUCTION

In any surveillance system, moving target detection from different environment such as indoor and outdoor with various challenges is very difficult task. Many researchers have worked on various challenges of object detection but still not developed a perfect algorithm. Camouflage occurs when moving object shares similar color with that of the background. Thus, as foreground pixels and background pixels appear same, it becomes very difficult task to separate moving target from background. This paper addresses dynamic background and camouflage problem. Basic methods of detection are background subtraction, frame differencing and optical flow. Then, came background modelling methods such as gaussian mixture model (GMM), kernel density estimation etc. and manual feature extraction.

Nowadays neural network has got utmost importance in research work especially in detecting moving objects. Neural networks are typically organized in different stages. These stages consist of a number of nodes(vertices). These nodes are linked to every other node. Every vertex contains a mathematical function. Input is applied through the input stage that passes through one or more inter-immediate stage where the operations are done.

An output stage gives the actual output. After convolution neural network (CNN) [1] came Region base convolutional neural network (RCNN) [2], fast RCNN [3], Faster RCNN [4] and Mask-RCNN [5] with little modifications in each stage for object detection. Our method uses Mask-RCNN along with some addition for accurate object detection. Mask-RCNN predicts object mask

or each target in parallel with classification and rectangle box. Histogram equalization concept to improve contrast [6]is applied on the target output followed by dilation for better results.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author

Archana Date*, Sinhgad College of Engineering, Pune, India. (Savitribai Phule Pune University), archanadate@gmail.com

Sanjeevani K. Shah, Smt. Kashibai Navale College of Engineering, Pune, India. (Savitribai Phule Pune University), san_shah@rediffmail.com

II. LITERATURE SURVEY

With increasing research work in moving object detection more challenges were discovered, hence it became an active research area. Lots of research has been done to develop new models to overcome problems such as camouflage, dynamic background, illumination changes, shadow, and occlusion.

Basic object detection algorithms are background subtraction [7], frame differencing [8], and optical flow [9]. Background subtraction gives poor results for dynamic background, a ghosting effect occur in case of frame differencing method whereas complexity increases in case of optical flow. In case of dynamic background, parametric pixelwise GMM [10] is used to model the multimodal background. But it is sensitive to noise. A model given in [11] deals with non-Gaussian processes is more robust to dynamic background then GMM. All these methods need prior knowledge of background.

Barnich et al. [12] proposed a Visual Background Extractor (Vibe) that does not depend upon any parameter and creates the background by summing earlier observed values for each pixel position. But not accurate for dynamic backgrounds and shadows. Pixel-Based Adaptive Segmented (PBAS) [13] approach updates background using recently calculated pixel values. The background model is upgraded every time so as to deal with gradual background changes. Some methods that are based on texture uses combination of different features [14] were used to detect object [15][16] (Mahajan 2019). This method fails for smooth surfaces and shadows. GLCM is also combine with HSV (hue, saturation and value) [17] to detect camouflage moving object. Nikos Paragios et al. [18] combined optical flow and color to detect the camouflage moving object due to different motion characteristics. Features in time and space domain [19] are also used to detect camouflage object. The robust object detection [20] is reached by the fusion of motion, color and contrast along with temporal and spatial features. Camouflage object detection is done based on Color and Intensity [21].

But is sensitive to changes in lights and also detects shadow as an object. Thus, to deal with camouflage object edge feature is also added along with color and intensity [22].

This algorithm handles shadow problem effectively in recognizing camouflaged object. A combination of parametric model and non-parametric [23] model is used to detect moving object. This method fails in case of occlusion. In [24] both discriminative (DM) modeling and camouflaged modeling (CM) are combined using Bayes theorem to detect camouflage moving object. But it is impossible to differentiate camouflage object from shadow areas.

A method is proposed in [25] that learns and organizes the background motion pattern (SOBS) and hence obtained good results in case of dynamic background. Further method combines RGBD and SOBS [26] detecting camouflage object. With the existing branch for segmentation, anabranch network [27] provides classification that predicts the probability of having camouflaged target in an image, which is combined with the segmentation branch to increase accuracy. Accuracy is less on some datasets. In [28], Muhammad et al. presented a CNN architecture to detect fire in a video surveillance application. the proposed deep learning method suffers from higher false alarm rate.

Thus, the existing algorithms are not perfect for object detection from dynamic background and also for handling camouflaged problem. Thus, in this article Mask RCNN with CLAHE is used to detect camouflage moving object as well as the moving object from dynamic background.

III. METHODOLOGY

This Object detection involves locating the positions of the objects and classifying each object into different classes. Thus, to draw the rectangular box around the object the object must be localized accurately. CNN is basically used for object detection and localization. There are four main versions of the technique, out of which the recent version is Mask R-CNN. Each version is elaborated below:

R-CNN: Selective search algorithm is used for localizing the object. After localization features of object are extracted via Conv Net, such as Alex Net, before a final set of object classifications are made with linear SVMs.

Fast R-CNN: After the regions of interest (ROI) given by deep CNN a layer is used to resize and fuse regions. After this classification is done through fully connected layer.

Faster R-CNN: Instead of using selective search a separate model is used that gives proposes regions that contains object. Due to this speed increases.

Mask R-CNN: To faster RCNN, it adds another overhead of adding mask for each detected target.

Thus, the methodology includes Mask RCNN followed by contrast limited adaptive histogram equalization for moving object detection from dynamic background and camouflage handling.

A. *Mask RCNN*

Mask RCNN is used to segment various targets in an input. Output of Mask RCNN is the classified target along with the rectangular boxes for the image or video input. Mask RCNN consists of two steps: First, it recommends the areas where there might be the presence of object based on the current frame or video. Second, it classifies the objects according to different classes, updates the rectangular box and creates a mask in pixel level of the object depending on the previous step. Both steps are connected to the strong feature pyramid network. ResNet 101 is used for feature extraction from the input in Mask R-CNN whereas Convnet is used by faster RCNN. So, the initial move is feature extraction and it is the input to the next step.

A model is applied on the output from the first step. This step generates the areas or maps that are likely to contain object of interest [4]. The outputs are of different shapes and dimensions. Thus, to convert them all to the same shape mathematical operations are performed on each feature grid. Then they are passed through fully connected network (FCN) which does prediction and classification.

Above steps are similar to that of faster RCNN. An additional overhead is added in Mask RCNN that generates the mask for each predicted object. The predicted regions are compared with certain threshold. If the area is greater than threshold than it is likely to contain the object of interest. Else avoid that particular region. This step is repeated for all the regions and then a set of regions is selected having area greater than threshold. After this a mask branch is attached to the existing architecture. The output is the targeted object with mask on it. The size of the mask varies with the object. Model which is trained previously on COCO dataset is used.

B. *Contrast limited adaptive histogram equalization (CLAHE):*

In histogram equalization image intensities (biran Abdullah) are adjusted to upgrade the clarity of the output. The traditional equalization method tries to equally distribute the intensity values in the whole image. Since in our proposed method a masked object is extracted and remaining part of the output is converted to black area, the conventional histogram equalization does not benefit image enhancement and will change the removed black area with other intensity values. Therefore, an adaptive histogram equalization method is used. In traditional methods (AHE) mapping of each pixel to an intensity value derived from the surrounding pixels is done by using a modification function. However, the AHE method has the disadvantage of noise over enhancement in cases of homogenous regions, or poor noise to signal ratio [29].

The Probability Distribution Function (PDF) of a digital image (Di) that has TN pixels with intensity values distributed in (M) values where n_{ik} is the count of intensity pixels at intensity level D_{ij} is calculated using equation (1). The progressive mass Function (PMF) is given by equation (2).

$$P_i\left(D_{ij}\right) = n_{ik} \ /TN \qquad (1)$$

$$C_i\,(D_{ik}) = \sum_{j=0}^{k} \mathrm{Pi\,(Dij\,)} \qquad (2)$$

Noise amplification is limited by limiting the slope of the PMF function. Here the histogram is cut at the set clip limit and intensities are evenly diffuse across each bin [38]. This is the difference between AHE and CLAHE. One of the main advantages of CLAHE in comparison with AHE is that it removes unwanted noise in the homogeneous areas of the image. Because of this CLAHE is used in object detection. The basic form of CLAHE algorithm enhances the contrast among small regions, i.e., tiles, of an image. The contrast of every block is individually upgraded to yield histogram that matches with the distribution type.

The adjacent blocks were consolidated by using bilinear interpolation and the picture gray scale qualities were adjusted by the altered histograms.

The essential calculations for image object utilizing CLAHE strategy is depicted underneath:

step 1: Divide the frame into non overlapping regions of size 8 * 8, every one of which relates to the set of 64 pixels.

step 2: Calculate histogram for every block.

Step 3: Set clip limit to crop the histogram.

step 4: Modify each histogram according to the clip limit. Transformation function for uniform distribution is given as [30].

$$P(x,y) = [Pg_{max} - Pg_{min}] * Ci(Dik) + Pg_{min} \qquad (3)$$

where

Pgmax=maximum pixel value

Pgmin=minimum pixel value

P (x, y) =computed pixel value

Ci (Dik)= progressive mass Function

For exponential distribution modification function for gray level is given as

$$P(x,y) = Pg_{min} - \left(\frac{1}{\alpha}\right) * \ln[1 - Ci(Dik)] \qquad (4)$$

where α = Clip limit

The PMF of Rayleigh function is

$$y = p(f(a/b)) = \int_0^x \frac{x}{b^2} e^{\left(\frac{-x^2}{2b^2}\right)}$$

(5)

Discriminative function accumulated image is then masked to hide some portions of an image and to reveal some portions. Here unwanted background is hided for better result and fewer complications. Every pixel has its intensity but the concept of image intensity does not exist. In term of grayscale image, pixel intensity is nothing but its brightness. Brightness increases with intensity and vice versa.

Setting of intensity level of the image is nothing but setting its radiance. Structuring elements such as kernel are used in morphological operations. Generally, these operations are performed on input but here these are applied on the target output. The most basic structural operations are dilation and erosion. Dilation fills the holes in an image by adding pixels, while erosion reduces thickness along the boundaries. Thus, to increases visibility of the object and fill empty portions of the targeted object dilation is performed on the output image.

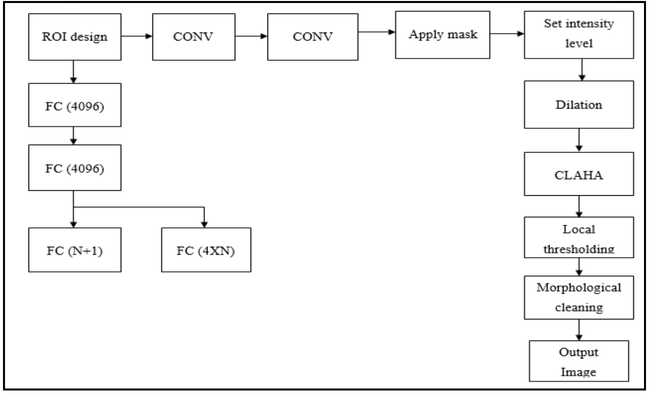


Fig 1: Proposed Method block diagram

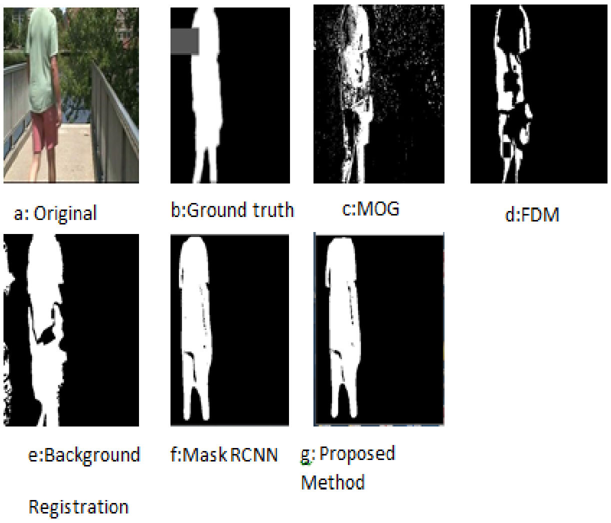
IV. IMPLEMENTATION RESULTS

Qualitative and Quantitative analysis is performed for both challenges for few scenarios from standard datasets. Sequences such as overpass sequence, water surface, curtain and fountain are used to analyze the performance of object detection from dynamic background. For camouflage challenge sequences are used from LASIESTA and CAM_UOW datasets.

For both dynamic background and camouflage, the results of proposed method were compared with the existing methods. The existing methods used for comparison for dynamic background are mixture of Gaussian, frame differencing method and background updating using background registration technique. Existing methods for camouflage are optical flow using HSV, histogram of oriented gradients and local binary pattern with SVM classifier and Mask RCNN without CLAHE. For all the standard datasets the results of the proposed method are excellent as compared to existing methods.

A. *Qualitative Analysis:*

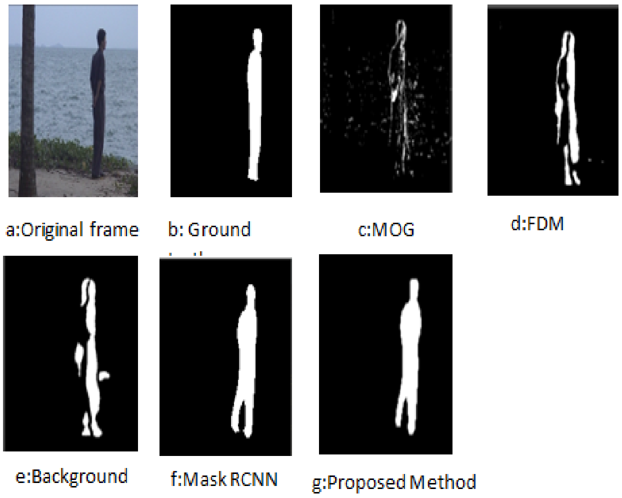
a. Overpass sequence


a: Original b:Ground truth c:MOG d:FDM

e:Background Registration f:Mask RCNN g: Proposed Method

Fig 2: Qualitative analysis of overpass sequence.

b. Water-surface sequence


a:Original frame b: Ground c:MOG d:FDM

e:Background f:Mask RCNN g:Proposed Method

Fig 3: Qualitative analysis of water surface sequence

c. Curtain Sequence

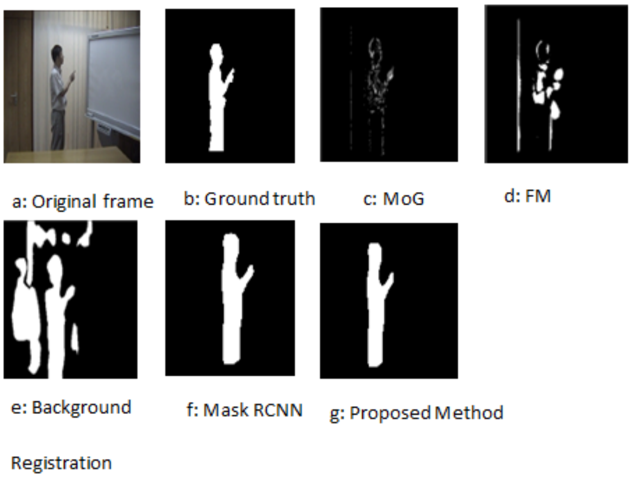


Fig 4: Qualitative analysis of curtain sequence

d. Fountain sequence

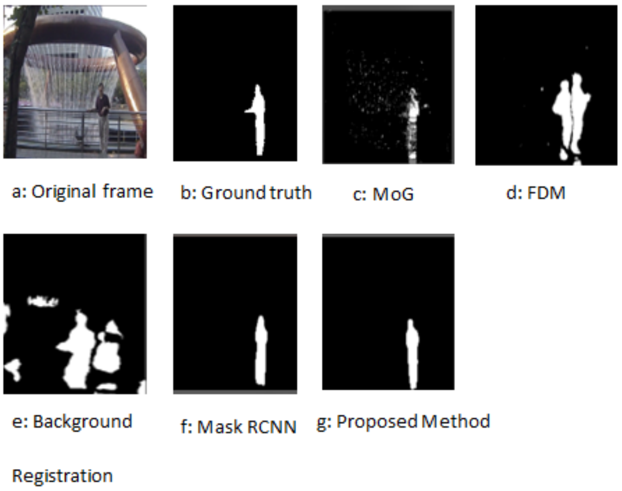


Fig 5: Qualitative analysis of fountain sequence.

e. LASIESTA-IC01 Sequence

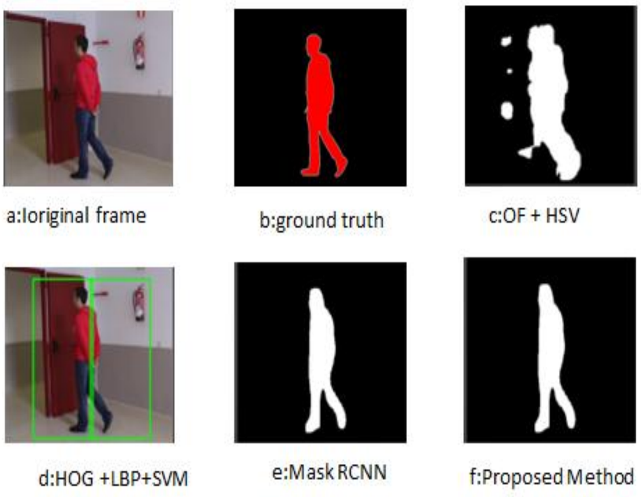


Fig 6: Qualitative analysis of IC01 sequence.

f. LASIESTA -IC02 Sequence

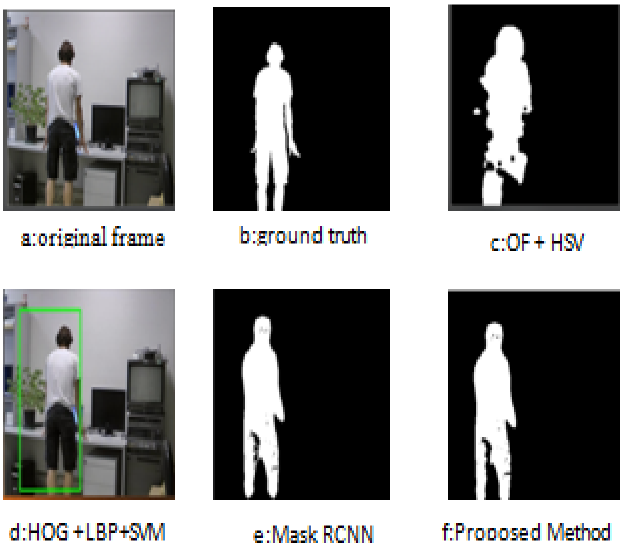


Fig 7: Qualitative analysis of IC02 sequence.

g. CAM_UOW video1

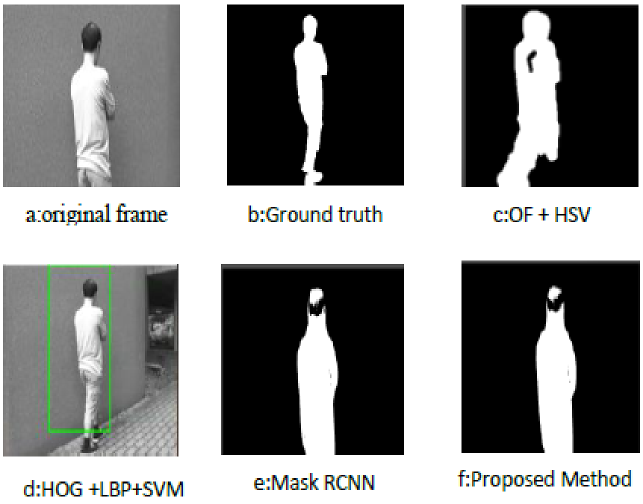


Fig 8: Qualitative analysis of CAM_UOW sequence.

h. CAM_UOW video2

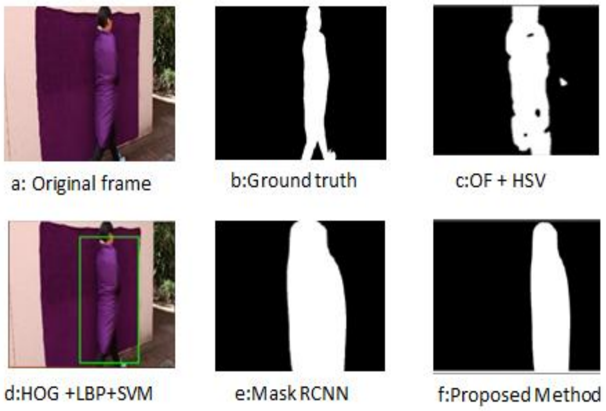


Fig 9: Qualitative analysis of CAM_UOW sequence.

B. *Quantitative Analysis*

Performance metrics considered were,

- Recall is known as detection rate.

$$Recall = \frac{Number\ of\ correctly\ identified\ pixel}{Number\ of\ foreground\ pixels\ in\ ground\ truth.}$$

- Precision is known as measure of exactness.

$$Precision = \frac{Number\ of\ correctly\ identified\ pixel}{Number\ of\ foreground\ pixels\ detected\ by\ algorithm}$$

- F measure is known as measure of merit.

$$F\ measure = \frac{2 * recall * precision}{recall + precision}$$

The table I shows quantitative analysis for dynamic background and table II shows quantitative analysis for camouflage moving object. As seen from the result the new method gives best result over the existing methods for dynamic background as well as camouflage.

Table I: Quantitative analysis for dynamic background

Sequences	**Curtain**			**Fountain**			**Water surface**			**Overpass**		
	R	P	F	R	P	F	R	P	F	R	P	F
FDM	99.2	42.9	59.9	76.37	33.0	46.1	93.8	32.0	47.8	78.2	89.3	83.2
BR	98.9	44.0	60.9	71.11	34.7	46.6	93.9	34.0	50.0	93.2	81.4	86.9
MOG	14.7	36.1	20.9	47.34	71.1	31.1	56.8	80.0	37.5	68.7	90.1	78.0
Mask RCNN	93.1	98.7	96.0	97.83	99.0	96.3	93.7	99.0	96.3	88.6	94.8	91.6
PROPOSED	93.6	98.7	96.1	99.41	99.7	99.5	93.7	99.1	96.3	88.9	94.6	91.7

The qualitative and quantitative results were obtained for dynamic background and camouflaged moving object. For Dynamic background four standard database sequences were used and for camouflage five standard database sequences were used. All the results of proposed method are compared and analyzed with existing algorithms. In both the challenges the results of proposed method both for qualitative and quantitative analysis are excellent as compared to existing methods.

Table II: Quantitative analysis for camouflage

Sequences	**HSV+OF**			**HOG +LBP+SVM**			**Mask RCNN**			**PROPOSED**		
	R	P	F	R	P	F	R	P	F	R	P	F
ICA01	77.37	72.54	74.87	49.00	1.00	66.00	93.40	98.88	96.06	93.40	99.05	97.17
ICA02	92.81	62.14	74.43	50.00	99.01	67.00	92.04	97.49	94.61	91.82	97.71	94.67
VIDEO1	68.84	50.65	58.72	52.43	97.14	68.10	90.96	98.47	94.56	91.5	97.92	94.60
VIDOE2	80.48	73.07	76.60	14.75	73.61	24.57	91.89	95.09	93.46	92.14	95.22	93.65

V. CONCLUSION

Methods such as RCNN, Fast RCNN and Faster RCNN classifies the object into different classes along with the bounding box. In case of Mask RCNN along with bounding box and classification, Mask is generated for each object. A neural network-based method, modified Mask RCNN using contrast limited adaptive histogram equalization is proposed and implemented to detect moving object. On the masked output contrast limited adaptive histogram equalization (CLAHE) is used to improve further results. Then by performing different morphological operations final object is detected. The method is tested on different standard datasets for detecting moving object from dynamic background and camouflaged handling. The proposed method is compared with existing algorithms for qualitative and quantitative measures. As seen from the results the proposed algorithm gives best results as compared to existing algorithms.

ACKNOWLEDGMENT

I express my sincere thanks to Dr. Sanjeevani S. Shah for her valuable advice, excellent guidance, support, encouragement and confidence towards me.

REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in NIPS, 2012.
2. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
3. R. Girshick. Fast R-CNN. In ICCV, 2015.
4. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
5. Kaiming He Georgia Gkioxari Piotr Doll´ar Ross Girshick Mask R-CNN. arXiv:1703.06870 v3 [cs.CV] 24 Jan 2018.
6. Zuiderveld K. Contrast limited adaptive histogram equalization. In: Graphics Gems IV. Academic Press Professional, Inc. 1994, pp. 474–485.
7. J. Heikkilä and O. Silvén, “A Real-Time System for Monitoring of Cyclists and Pedestrians,” Proc. of Second IEEE Workshop on Visual Surveillance, pages 74–81, JUNE 1999.
8. V. Sharma, N. Nain, and T. Badal, “A Survey on Moving Object Detection Methods in Video Surveillance,” vol. 2, no. 1, pp. 209–218, 2015.
9. Y. Zinbi and Y. Chahir, “ Moving object Segmentation using optical flow with active contour model”. IEEE Conference on ICTTA, pp.1-5, 2008.
10. Stauffer and E. Grimsson. Adaptive background mixture models for real-time tracking. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 1999, pages 246–252, 1999.
11. T. Haines and T. Xiang. Background subtraction with Dirichlet processes. European Conference on Computer Vision, ECCV 2012, October 2012.
12. O. Barnich and M. Van Droogenbroeck. ViBe: a powerful random technique to estimate the background in video sequences. International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, pages 945–948, April 2009.
13. M. Hofmann, P. Tiefenbacher, and G. Rigoll. Background segmentation with feedback: The pixel-based adaptive segmented. IEEE Workshop on Change Detection, CVPR 2012, June 2012.
14. Nagappa U. Bharati and P Nagabhusan, “Camouflage Defect Identification: A Novel Approach”, 9th International Conference on Information Technology (ICIT'06), 2006.
15. C. H. Yeh, C. Y. Lin, K. Muchtar and L. W. Kang, **“**Real-time background modeling based on a multi-level texture description**”** *Inf. Sci.*, vol. 269, pp. 106-127, 2014.
16. Teck Wee Chua, Yue Wang, Karianto Leman, "Adaptive texturecolorbased background subtraction for video surveillance", *19th IEEE International Conference onImageProcessing (ICIP)*, pp. 49-52, 2012.
17. R. E. Ch. Kavitha, B. Prabhakara Rao, A. Govardhan, “An Efficient Content Based Image Retrieval Using Color and Texture of Image Sub blocks”, International Journal of Engineering Science and Technology (IJEST) Vol. 3, No. 2, Feb 2011.
18. A. Mittal, N. Paragios, “Motion-based background subtraction using adaptive kernel density estimation”: IEEE Conference on Computer Vision and Pattern Recognition. 302–309, 2004.
19. L. Li, W. Huang, I. Gu, Q. Tian, “Statistical modeling of complex backgrounds for foreground object detection” IEEE Trans. Image Process. 1459–1472, 2004.
20. A. Criminisi, G. Cross, A. Blake, V. Kolmogorov, “Bilayer segmentation of live video” in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 53–60, 2006.
21. I. Huerta, D. Rowe, M. Mozerov, and J. Gonzalez, “Improving Background Subtraction Based on a Casuistry of Color-Motion Segmentation Problems” IbPRIA 07, Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part II, pp. 475 – 482, 2007.
22. P. Siricharon, S. Aramvith, T.H. Chalidabhongse and S. Siddhichai. “Robust Outdoor Human Segmentation based on Color- based Statistics Approach and Edge Combination”, 2010.
23. Zhou Liu, Kaiqi Huang and Tieniu Tan “Foreground Object Detection Using Top-down Information Based on EM Framework”, IEEE Transactions on Image Processing, 2011.
24. X. Zhang, C. Zhu, S. Wang, Y. Liu, and M. Ye, "A Bayesian Approach for Camouflaged moving Object Detection," IEEE transactions on circuits and systems for video technology, vol. 1, no. d, pp. 1–13, 2016.
25. Maddalena L, Petrosino A (2008) A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans Image Process 17(7):1168–1177.
26. Maddalena L, Petrosino A (2018) Background subtraction for moving object detection in RGBD data: a survey. J Image 4(5).
27. Trung-NghiaLe , Tam V. Nguyen, ZhongliangNie , Minh-Triet Tran, Akihiro Sugimoto “Anabranch network for camouflaged object segmentation”, Computer Vision and Image Understanding 184 (2019) 45–56,2019.
28. Muhammad K, Ahmad J, Mehmood I, Rho S, Baik SW (2018) Convolutional Neural Networks Based Fire Detection in Surveillance Videos. IEEE Access 6:18174–18183.
29. K. Zuiderveld “Contrast Limited Adaptive Histogram Equalization.” Graphic Gems IV. San Diego: Academic Press Professional, 1994. 474–485.
30. Muhammad SuzuriHitam,. Wan NuralJawahirHj Wan Yussof, Ezmahamrul Afreen Awalludin, Zainuddin Bachok, “Mixture Contrast Limited Adaptive Histogram Equalization for Underwater Image Enhancement”, IEEE international conf. 2013.

AUTHORS PROFILE

Archana R. Date obtained her BE (Industrial Electronics) degree from Pune University, Pune in 1998 and MTech (Electronics) degree from Bharati Vidyapeeth deemed university, Pune in 2011. Presently a research scholar at Sinhgad college of engineering, Pune.Published papers in national, international conferences and journals.

Dr. Sanjeevani K. Shah obtained her Ph.D. (E&TC) from Savitribai Phule Pune University, Pune in 2012. Worked in Philips India Ltd. for three Years. Thereafter has thirty-three years of teaching experience. Presently working as Head of Post Graduate department E&TC in STES’s SKN College of engineering. Published books on Industrial Electronics, Communication, applied electronics and has published papers in national, International conferences and journals. The research paper ‘Microwave drying of biomass: A remedy to environmental hazards due to uneven combustion of biomass’ published in Int. Journal of IEI got Nawab Zain Yar Jung Bahadur Memorial Prize in the year 2012.