International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Mask Region Based Convolution Neural Network (R-CNN) based Smart System for Anomaly Detection in Pedestrian Walkways

A. Nirmala, S. Arivalagan, P. Sudhakar

Abstract*: Recently, anomaly detection becomes a fascinating research application which usually raises an alarm in scenarios where the event varies from the actual event. Anomaly detection can be treated as a coarse-level video understanding problem that determines the existence of anomalies from habitual events. This paper introduces a new anomaly detection model by the use of Mask region based convolution neural network (R-CNN). The application of mask in the detection process helps to precisely identify the presence of anomalies in the scene. The effectiveness of the Mask R-CNN based anomaly detection model is validated against UCSD anomaly detection dataset. An extensive quantitative and experimental outcome evidently shows the superior nature of the presented model over the compared methods in a significant manner.

***Keywords*: Anomaly detection; Deep learning; RCNN; Object identification.**

I. INTRODUCTION

In public places such as traffic signals, shopping malls, roads, banks, etc, surveillance cameras are installed presently to improve public security level. It is a complicated procedure to observe video consequently at a faster rate. For monitoring, it tends to inadequate surveillance cameras utilization and needs to presence of human. In video surveillance, the major complexity lies in the process of anomaly detection such as thefts, any illegal actions, accidents, or crime. While comparing to common events, anomaly actions will not happen often. For anomaly detection, time and human resources are reduced by the use of smart computer vision techniques [1]. Raising an alarm at such circumstances is the major purpose of 'anomaly detection technique' wherever the typically action deviate from the original action. Therefore, detection of anomaly might be seen as problem of video understanding at coarse-level that recognizes the anomalies occurrence out of the actual action. By employing classification techniques, it might be categorized into any certain event while the anomalies are detected successfully. To recognize the certain anomalies, the model of anomaly detection, for example, traffic accident and violence detector has been introduced. To recognize the anomalies, it is noted that the solutions might not be in discriminate and they are restricted to limited usage [2]. Few real time anomalies are distinct and composite, when it is necessary to denote each anomaly. It is desirable to build the method of anomaly detection which is not based on any data over anomalies. Without more supervision, anomaly detection might be performed at the same time. The techniques of Sparse coding-based are assumed as representative methods that succeeds the traditional techniques of anomaly detection. Only a primary or short video portion held the actual activities in order to the methods where those parts might be used to build the actual event dictionary. For anomaly detection, the main idea is the anomaly should be reconstructed accurately out of the actual action dictionaries. The surveillance cameras captured videos comprise changing actions rapidly. For various actual actions, the traditional techniques created huge false detection.

For pedestrian detection, different techniques might be built that adopts bounding boxes for every pedestrian exist within the image [3, 4]. Over the computer vision group, it derives higher interest as an important component for different applications that are human-oriented such as automatic traffic signalling, driverless cars, person identification, etc., These techniques fails to solve a main complexity of scaling that stay unresolved in addition it might impact mainly the technique of pedestrian detection results in common scenes. Previous work done before gives to scaling issue resolving depending on two dimensions. To improve the capability of scale-invariance capability primarily and the data of brute-force was augmented. Subsequently, in each sample with various sizes, single model with multi-scale filters were employed. However, it is a complex procedure because of the small and large-sized intra-class variance samples majorly-various feature responses with distinct model. To make use of the varying case features with various scales, divide-and-conquer ideology is employed by the author to resolve the crucial scale variance problem [5, 6].Few deep learning [7] based techniques of anomaly detection were built recently. Convolution Neural Network (CNN) was employed primarily that categorizes the object present in an area as normal or anomaly.

Manuscript published on 30 September 2019
* Correspondence Author
A. Nirmala*, Research scholar, Department of computer science Annamalai University, Chidambaram, India. Email: sjcnirmala@gmail.com
Dr. S. Arivalagan, Assistant Professor, Department of computer science and engineering, Annamalai University Chidambaram, India. Email: arivucseau@gmail.com
Dr. P. Sudhakar, Assistant Professor, Department of computer science and engineering, Annamalai University Chidambaram, India. Email: kar.sudha@gmail.com

The images also suffer from few problems like objects aspect ratios and various spatial locations in the image. There exists a requirement to select a count of areas that gives higher computation complexity [8].

To recognize the occurrences, YOLO, region-based convolution neural network (R-CNN), etc were given. To solve the problems, novel methods were used in selecting the higher count of areas and this method employed selective search technique for deriving 2000 areas from the image that were called as region proposals. Only 2000 areas were employed, in spite of categorizing higher counts of areas. These 2000 regions were created by the technique of selective search. It faces below complexities even though it minimized the CNN computation complexity slightly. For each image, the time takes to train the network is high in classification of 2000 area proposals. For each test image, to execute in real time applications, it was not probable as it needs approximately 47s. The technique of selective search is a fixed technique. There is no requirement for learning procedure which leads to the bad tentative region proposals creation. Through the similar researches, the constraint was solved by the development of a rapid object detection method known as R-CNN. Excluding the offer of providing an input image to generate convolutional feature map, it is same as R-CNN in spite of offering area proposals towards CNN. It is called as Fast R-CNN where it is not needed to offer 2000 region proposals towards CNN every time. Degraded performance was attained through Fast R-CNN; however, selective search technique was employed. A novel object detection method was given by [9] through reducing the selective search technique usage wherever the knowledge is acquired by the network out of the area proposals itself. For the region proposal prediction, the distinct network usage is the difference among the fast R-CNN and faster R-CNN wherever the reshaping is done by RoI pooling layer. In the proposed area, the images were categorized and for each bounding box, offset rate is predicted.

For anomaly detection, Mask RCNN is introduced in this paper for pedestrian walkways. For faster identification, region-based proposals are employed by the projected technique and with masking procedure. Over the pedestrian walkways, the projected model of Mask R-CNN recognizes the anomaly presence such as skaters, bikers, vehicles, and so on. By employing various video series, Mask R-CNN model was verified from the UCSD anomaly detection dataset. A comparative results analysis takes place with Fast R-CNN, Mixtures of Probabilistic Principal Component Analysers (MPPCA), Minimization of Drive Testing (MDT), and Social Force (SF) the outcomes were compared. The simulation results give the efficient detection performance of the presented model over the entire set of given test images.

The rest of the portions are planned as follows. Section 2 presents the Mask R-CNN model for anomaly detection. Section 3 investigates the experimental results and Section 4 concludes the paper.

II. RELATED WORK

Detection of anomaly is a highly complex and challenging problem in computer vision [10, 11]. For video surveillance application, to recognize the violent or aggressive behaviour in videos, many attempts were done [12, 13]. A technique of people violent behaviour detection is presented by the individual motion of limbs orientation. For the detection of aggressive actions, the researcher employed video and audio data in surveillance videos. In crowded videos, to detect violence, violent flow descriptors were employed. For the classification of violent and non-violent videos [14], an effective heuristic dependent technique was projected. To follow the actual people activity and to categorize the abnormal actions as anomaly in under violent and non-violent patterns discriminations, a tracking method is proposed by [15], to recognize the actual activity deviations. To acquire knowledge over preventing tracking and global motion, Hidden Markov Model (HMM) over local spatiotemporal volumes, context-driven technique, SF model, Dynamic texture mixtures model, and topic modelling are developed. The techniques with normal behaviour, learnt from training videos and low probability patterns are detected as anomalies. In computer vision problems, from dictionaries to actual behaviour learning, sparse representation is used by [16]. On pattern testing, anomaly behaviour was found which contains large reconstructing bugs. By employing deep learning [17], numerous techniques were proposed for the video action classification. Deriving annotation is laborious and complex to train in case of videos.

Cognition emulation in human being decision making in simulation had been studied in dynamic fields and aids for individual role in simulation. Data-driven techniques use machine learning for mapping agent stimuli to activities while comparing towards cognitive techniques. An agent may outcome out of the superior match databases, while the techniques search fitting over extensive range. Database clustering techniques searches the superior possibility may lead to different same stimuli activities. Traditional techniques might be proposed by subjective observation [19], though the crowd is analysed. CNN had been used in general object identification successfully. The present researchers focused on improving the pedestrian detection performance using the techniques of deep learning [5,6]. For pedestrian detection, sparse convolutional coding is used by [20] for CNN pre-training depending on the unsupervised methods. The improved pedestrian detection is merged with semantic task [21]. Numerous techniques were projected to improve the scale-invariance of CNN. The derived CNN activation was used for local patch over the three various scales [22]. By order-less pooling performance, combined patch characteristic was generated at each level distinctly. At different scaling levels, input patterns are recognized through [23] in a column count and merged the feature in top-layer which map the full columns. Many anomaly detection techniques available in the literature were time taking and it fails in detecting anomalies efficiently. This study focuses on building an effective and faster technique of anomaly detection by employing Mask R-CNN model.

III. PROPOSED METHOD

A. Overview

To recognize the target objects existence robustly and precisely, Mask R-CNN is employed by the projected MOT technique. Fig. 1 shows the procedure involved in the projected model. The applied video input is categorized into frame set primarily. And, to enhance the tracking results, the procedure of feature extraction is carried out.

Towards each object, the procedure of feature extraction might allocate class labels with the structured mask. Towards model training, Mask-RCNN is used. Novel test video series might be used for monitoring multiple objects existence within the frame when the training of the model is done. Towards a frame set, the video series would be divided to model testing as same as the training procedure. With an applied label, a mask might be created when frames are tested.

B. Mask R-CNN

Mask R-CNN [24] is theoretically simple. It defined that two outputs are comprised by the Faster R-CNN for each candidate object that are class label and bounding-box offset. The consequent branch is included which offers object mask. Mask R-CNN is common and intuitive concept. The additional mask output is distinct which requires extraction of spatial object layout that is delicate from the class and box outputs. Pixel-to-pixel alignment is the major Mask R-CNN components that are not present in the Faster R-CNN as demonstrated in Fig. 2. The same two-step procedure is used by Mask R-CNN with similar primary stages. For each RoI, Mask R-CNN offers a binary mask with class prediction and box offset in the second phase. With the modern system, this is highly contrasting whereas the classification depends on predictions of mask.

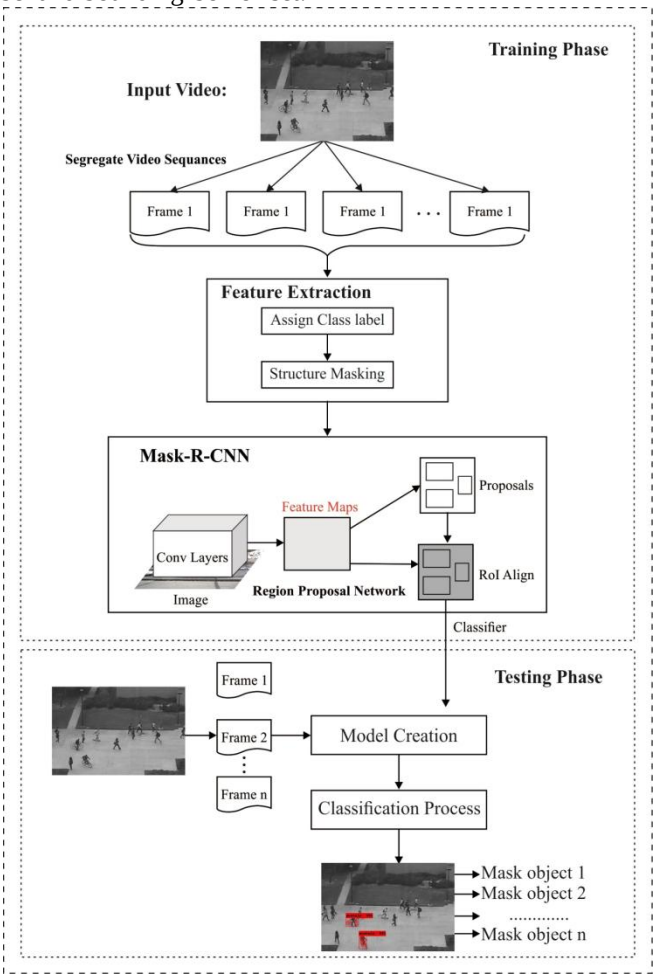


Fig. 1. Overall process of the presented Mask R-CNN based anomaly detection model

L_{mask} I described over k-th mask for RoI given ground-truth class k. RoI L=L_cls+L_box+L_mask is the multi-task loss that are discussed while training over every samples. L_box denotes the bounding-box and L_cls denotes the classification loss that is identical. For every RoI, mask branch is comprised by Km^2dimensional output that encodes the m × m resolution K binary masks for every K class. Per-pixel sigmoid is employed and describes loss of average binary cross-entropy as L_{mask}

For every class, the definition of L_{mask} allows to create masks for every class with class competition. It is based on dedicated classification branch for prediction of class label that is used to select the output mask. This decouples mask and class prediction. This differs from common works when employing FCNs for semantic segmentation that uses per-pixel softmax and multinomial cross-entropy. With a per-pixel sigmoid and binary loss, the masks on classes compete over one another. For gaining better segmentation outputs, we show that this is major thing.

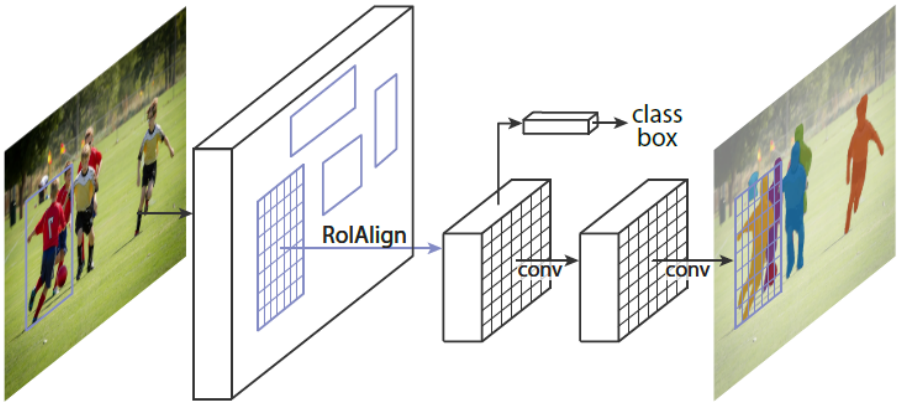


Fig. 2. Mask R-CNN

The loss function to define an image is defined as follows:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}$$

(t_i, t_i^*)

where i is the index of an anchor in a mini-batch and p_i is the predicted possibility of an anchor i might be an object. The ground-truth label p_i^* is 1 when the anchor is positive or it is kept as 0 when the anchor is negative. t_i is a vector representing 4 parameterized coordinate points of the predicted bounding box, t_i^* is a ground-truth box connected to an anchor which is positive. The classification loss L_{cls} is log over two classes. For the regression loss, $L_{box}(t_i, t_i^*) = R(t_i - t_i^*)$, where R is the robust loss function (smooth L1). The element $p_i^* L_{reg}$ represents the regression loss which is the inactive state for $p_i^* = 1$ and in the inactive state for $p_i^* = 0$. The output of the cls and reg layers has $\{p_i\}\ and\ \{t_i\}$. To represent a bounding box regression, the parameters are assumed as provided in Eqs. (2)-(5).

$$t_x = (x - x_a)/w_a, \quad t_y = (y - y_a)/h_a \qquad (2)$$

$$t_w = log\,(w)/w_a, \quad t_h = log(h)/h_a \qquad (3)$$

$$t_x^* = (x^* - x_a)/w_a, \quad t_x^* = (y^* - y_a)/h_a \qquad (4)$$

$$t_w^* = log(w^*)/w_a, \quad t_h^* = log\,(h)/h_a \qquad (5)$$

where x, y, w and h represents the box's centre coordinates and its width and height. Variables x, x_a and x^* denotes the predicted box, anchor box and ground truth box.

C. Mask Representation

By the mask, an input object's spatial layout is encoded. Hence, not like class labels or box offsets that are combined into shorter output vectors, deriving spatial mask framework may be examined by the pixel-to-pixel way which offers by convolutions. With m × m as dimension, the mask is predicted by using FCN. It allows each layer of mask branch into vector representation without combing it and to handle object spatial layout as m × m that absent in spatial dimensions. Few parameters are required by fully convolutional representation in spite of past method which employs layers of fc for mask prediction and it is greatly precise. Pixel-to-pixel behaviour is required by RoI features that are smaller feature maps which requires to be managed to per-pixel spatial association conservation explicitly. In mask prediction, this motivated to model a RoIAlign layer below which play as a lead role.

D. RoIAlign

For each RoI, for a small feature map RoI Pool is a main operation. RoI Pool quantizes primarily the floating-number RoI to the feature map discrete granularity which quantifies RoI and partitions into spatial bins which are itself quantized and final feature rates are masked by each bin that are merged. For example, Quantization is performed by computing $[\frac{x}{16}]$ on a coordinate x that is continuous, where 16 denotes the feature map stride and rounding is represented by $[\cdot]$. Between the RoI and derived features, this process introduces misalignments. Over the prediction of pixel-accurate masks, it contains the high negative impact.

To overcome this, we present RoIAlign layer that removes the harsh quantization of RoIPool and that align perfectly the extracted input features. Bilinear interpolation is employed to compute the accurate rate of input features in each RoI bin and merges the outcomes. The outputs are not sensible to the accurate sampling locations. Higher improvements are given by RoIAlign. The proposed operations of RoIWarp are compared. RoIWarp is the alignment issue as quantizing of RoI is similar to RoIPool. Bilinear resampling is used by RoIWarp: For each RoIPool, it denotes the crucial alignment role.

Backpropagation send the derivatives via the RoI pooling layer. To clarify this, it is assumed that only one image is present in a mini-batch (N = 1), even though the extension to N > 1 is direct due to the fact that the forward pass handles every image in an independent way. Let $x_i \in R$ is the i-th activation input to the RoI pooling layer and let y_{rj} be the layer's j-th output from the e rth RoI. The RoI pooling layer computes

$$y_{rj} x_i * (r,j) = \arg \max_{i' \in R(i,j)} x_{i'} \cdot R(i,j) \qquad (6)$$

is the index set of inputs in the sub-window over which the output unit y_{rj} max pools. A single x_i can be allocated to numerous outputs y_{rj}.

The RoI pooling layer's backwards function determines the partial derivative of the loss function based on every input variable x_i using the argmax switches as given below:

$$\frac{\partial L}{\partial x_i} = \sum_r \sum_j [i = i * (r,j)] \frac{\partial L}{\partial y_{rj}} \qquad (7)$$

For every mini-batch RoI r and for every pooling output unit y_{rj}, the partial derivative $\frac{\partial L}{\partial y_{rj}}$ is derived when i is the argmax chosen for y_{rj} by max pooling.

In back-propagation, the partial derivatives $\frac{\partial L}{\partial y_{rj}}$ are determined earlier through the backward function of the layer on top of the RoI pooling layer.

E. Network Architecture

We commence the multi-framework Mask R-CNN to establish the generality technique. For clarity, we establish: (1) the structure of convolutional backbone for feature extraction (2) To each RoI, mask prediction and network head are employed independently that are bounding-box recognition. We represent the structure of backbone using nomenclature network-depth-features. Over the layer depth from 50 to 101, we analyze the ResNet and ResNeXt networks. With the extracted features of ResNets, the implementation of original Mask R-CNN is done at the 4-th phase. The backbone with ResNet-50 is represented by ResNet-50-C4. Other effective technique employed is Feature Pyramid Network (FPN). From a single-scale input with lateral associations, top-down structure is used by FPN to build network feature pyramid. From different levels of feature pyramid, RoI features are derived through Faster R-CNN to scale with FPN backbone. As similar to vanilla ResNet, the other techniques are also identical. For network head with following structure that present in a previous work, we employ fully convolutional mask prediction branch. The fifth stage of ResNet is involved by head over ResNet-C4 backbone that is intensive computationally. For FPN, res5 is involved by the backbone and allows efficient head that uses few filters. It has the ability to improve the performance and it is a standard structure containing complicated designs.

F. Implementation Details

We employ hyper-parameters after the work of Fast/Faster R-CNN. These were performed for object detection and the sample segmentation system is robust towards it.

Training: A RoI is considered to be positive as same as Fast R-CNN, while IoU with real-truth negative box otherwise with 0.5 at least. L_{mask} denotes the mask loss which is denoted as positive RoIs. Mask target is the intersection between ground-truth mask and RoI. We employ training that is image-centric. By keeping in mind, the size as 800 pixels, the images are restructured. Each mini-batch contains two images for every GPU and each image contains N sampled RoIs with proportion of 1:3 of positive to negatives. N is 64 for C4 backbone and N is 512 for FPN. By assuming learning rate as 0.02, we train 8 GPUs for 160k iterations that are minimized by ten iterations at 120k. We use 0.0001 as weight decay and 0.9 as momentum. For every GPU, we train one image with ResNeXt and same iteration counts with starting rate of 0.01. RPN anchors span 5 scales and 3 aspect ratios. For suitable ablation, RPN is trained separately and it does not spread the Mask R-CNN features unless noted. For every entry, RPN and R-CNN contain same backbones as they might be distributed.

Inference: During testing time, the proposal count is 300 for C4 backbone and 1000 for FPN. We implement branch subsequent box and box prediction based on the proposals. Mask branch is employed to the huge number of 100 detection boxes scoring. It improves the precision and boost up the inference, though it modifies from parallel computing. By the mask branch, K masks for each RoI might be predicted; we use k-th mask, whereas the predicted class is denoted by k by the classification branch. Floating-number mask output is re-built by 0.5 to the RoI size as threshold. On the top of 100 detection boxes, little overhead is offered by Mask R-CNN to the Faster R-CNN counterpart.

IV. Performance Evaluation

Over various image series, we examine the Mask R-CNN model performance effectively. The used parameter is of learning rate 0.02, 8 as size of batch, 10000 as step size or epoch, 600 as minimum dimension, 0.7 as score threshold and 1024 as maximum dimension. MDT, SF, Fast R-CNN, and MPPCA are employed for comparison purposes. In the below subsections, the measures, used dataset and outcomes are discussed.

A. Dataset

Anomaly Detection Dataset from UCSD is employed for validation [25]. A set of images that are taken from stable camera is contained in the dataset of UCSD Anomaly Detection at the pedestrian walkways that is elevation overlooking. Towards over-crowded from sparse, the crowd density within the walkway is not a stable range. The video comprises abnormal actions, pedestrians or anomalies in actual cases that involve the non-pedestrian entities movement in walkways. Skaters, small carts, bikers, vehicles, bikers, and people walking or in grass are the present anomalies that surrounds it. The dataset details are provided in table 1.

Table 1 Dataset Description

Dataset	Testbed	Frames	Time (s)
UCSDped2	Test004	180	6

B. Result analysis

A quantitative analysis of the presented Mask R-CNN model is shown in Fig. 3. It is absolute from figure that the projected model of MASK R-CNN detects successfully the biker and cart with the detection rate of 99% and 97% correspondingly. The biker is detected by Fast R-CNN accurately as shown in Fig 3; however, it fails in detecting the cart and attains 55% as low accuracy.

(a) Social Force

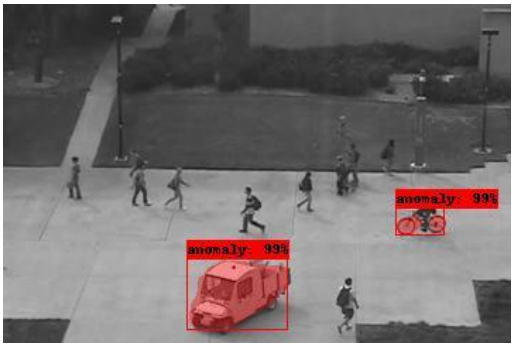
(b) Faster R-CNN

(c) Fast R-CNN

(d) Spatial MDT

(e) Temporal MDT

(f) MPPCA



(g) Mask R-CNN

Fig. 3. Quantitative analysis of various methods on UCSD Anomaly dataset

Table 2 Accuracy of anomalies per frame identified in Test 004

Frame Number	Anomaly 1	Anomaly 2
040	99	-
042	99	-
046	99	-
051	99	-
075	99	-
106	99	-
123	97	-
135	97	97
136	97	97
137	97	97
149	99	97
158	97	97
177	99	97
178	99	98
180	97	99

In addition, worst performance is demonstrated by MPPCA and SF when comparing with the other techniques as it detects wrongly the two anomalies. It wrongly detects pedestrians as anomaly. On the other hand, the spatial and temporal MDT attempts to handle well and detects one anomaly efficiently. At the same time, anomaly cart is detected correctly by spatial MDT and the biker alone is detected through temporal MDT. However, the proposed method creates a mask over the anomaly object and gives label correctly. Hence, enhanced performance is shown by proposed method on all the applied images. Fig. 4 and Table 2 show the detection rate of the presented model on various frames.

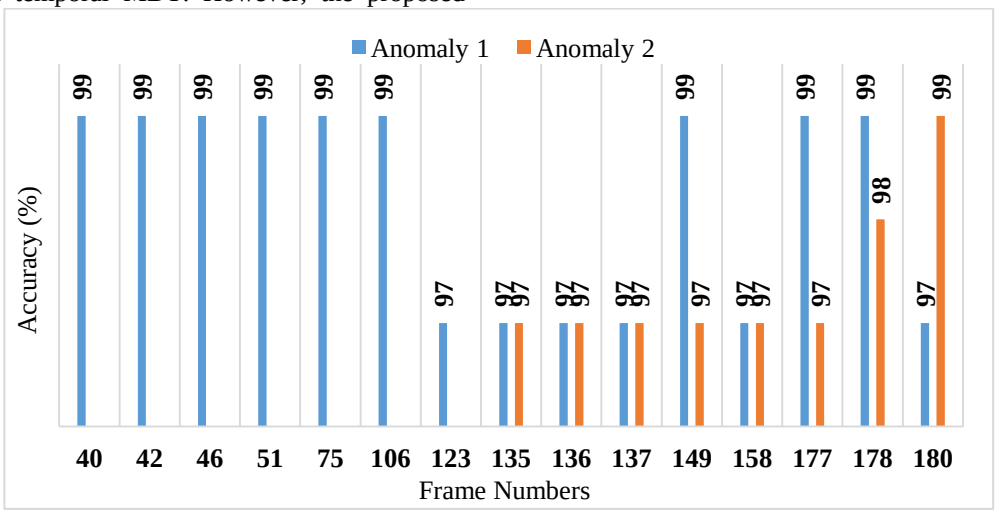


Fig. 4. Accuracy of Anomaly 1 and Anomaly 2 for Test004 Dataset

Table 3 Anomaly Detection accuracy comparisons of various methods for Test 004 Dataset

Frame number	Proposed	Faster R-CNN	Fast R-CNN	MDT	MPPCA	SF
040	99.00	94.00	81.92	76.83	74.37	52.36
042	99.00	94.00	82.35	76.63	78.23	63.94
046	99.00	95.00	85.33	75.24	76.84	53.58
051	99.00	96.00	79.34	89.83	75.90	60.13
075	99.00	99.00	78.34	82.73	75.21	52.37
106	99.00	99.00	91.23	83.41	72.27	51.34
123	97.00	97.00	91.28	87.92	71.39	57.49
135	97.00	97.50	92.37	80.62	77.46	53.56
136	97.00	98.00	94.61	83.91	71.26	63.23
137	97.00	98.50	91.72	85.63	75.37	57.94
149	98.00	99.00	83.90	78.56	70.94	61.32
158	97.00	99.00	77.14	78.32	71.57	54.43
177	98.00	99.00	79.34	75.34	76.95	52.21
178	98.50	98.50	82.37	76.45	80.23	51.35
180	98.00	99.00	85.33	85.23	70.38	60.47

Fig. 5 and Table 3 demonstrate the comparative results of different frames and the accuracy values are given below.

On the whole, the proposed method attains maximum accuracy in all the cases for the given dataset, and the SF attains worst performance. For instance, over the applied dataset, for frame number 040, the worst performance is demonstrated by SF of 52.36%, whereas MDT and MPPCA do not attain considerable performances. Fast R-CNN demonstrates 81.92% as accuracy rate. Faster R-CNN attains enhanced performance over others by attaining 94% as accuracy rate whereas the proposed method attains superior performance of 99%.

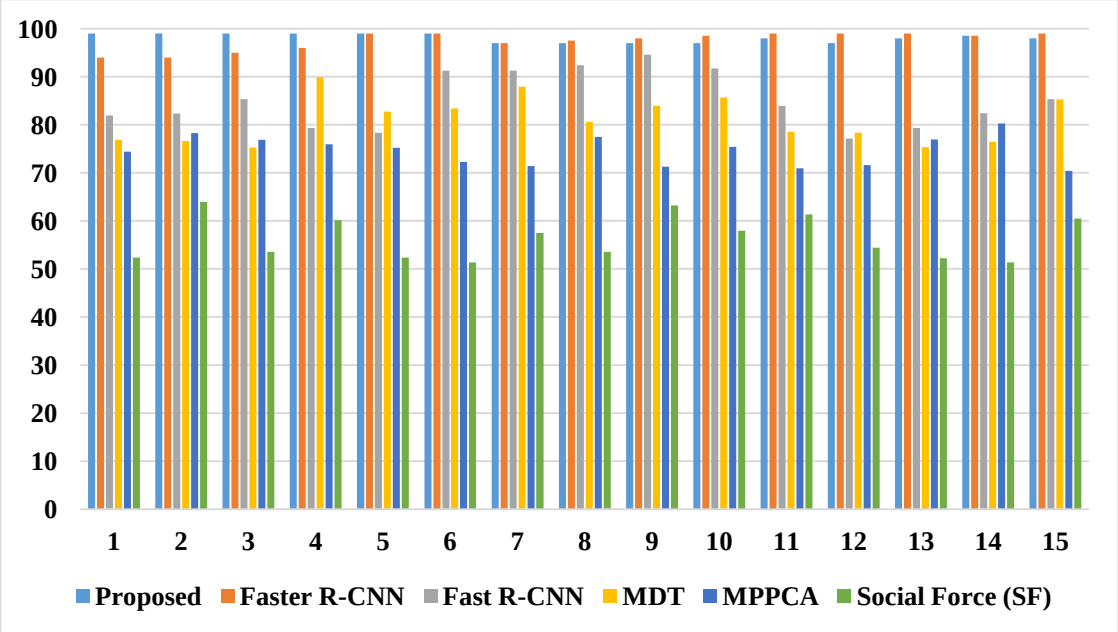


Fig. 5. Comparative results of various detection methods concerning accuracy for Test004

The average accuracy of various methods is given in Fig. 6. With all the compared methods, MPPCA is the worst method that shows the accuracy of 74.558%. The methods like MDT and Fast R-CNN does not show superior performances by attaining the MDT and Fast R-CNN rate of 81.11% and 85.105%. Even though, the Faster R-CNN and attains superior performance of 97.5, it fails to succeed the proposed method that achieves the 98.1%. The proposed method achieves the superior accuracy rate among all the other methods.

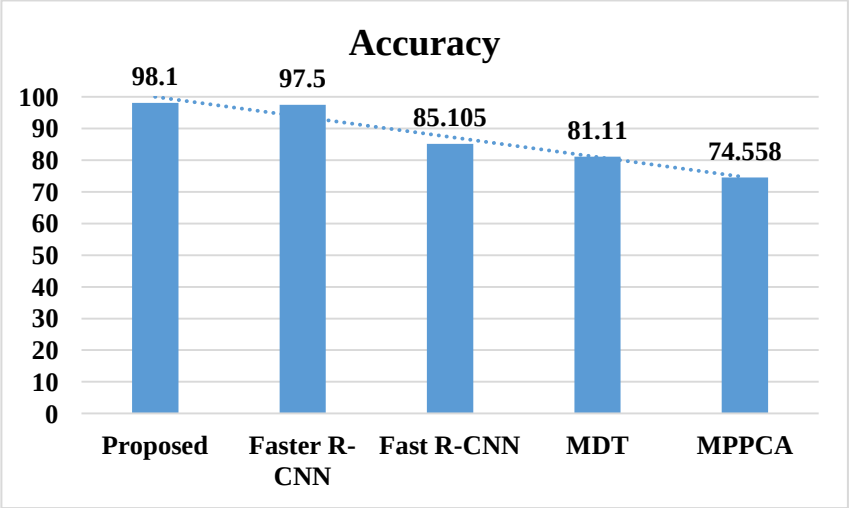


Fig. 6. Comparison of different methods in terms of average detection accuracy

V. CONCLUSION

In the applications of computer vision, detection of anomaly in surveillance videos is a composite task. To control the anomaly or unusual behavior, the technique of automated anomaly detection might be used as a smart model in the video sequences. In the pedestrian walkways, to recognize the anomaly existence, we presented an effective method of anomaly detection known as Mask R-CNN model in this paper. The projected system is efficient and simple and might recognize the anomalies at a faster rate. The feature of compatibility of various sized anomaly creates it efficient and tend to improved performance in detection. With the compared techniques, the outcomes denoted efficient performance of proposed method with the wide experimentation. The higher average accuracy is demonstrated by proposed Mask R-CNN model of 98.1 wherever MPPCA, MDT, Fast R-CNN and Faster-R-CNN achieve the accuracy rates of 74.558%, 81.11%, 85.105% and 97.5% respectively. As a part of future work, the presented model can be employed in public places.

REFERENCES

1. Hu, W., Tan, T., Wang, L. and Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(3), pp.334-352.
2. Mi, Z., Shang, L. and Xue, B., 2018, December. Multi-Dimensional Optical Flow Embedded Genetic Programming for Anomaly Detection in Crowded Scenes. In International Conference on Neural Information Processing (pp. 486-497). Springer, Cham.
3. Wang, T., Qiao, M., Lin, Z., Li, C., Snoussi, H., Liu, Z. and Choi, C., 2018. Generative Neural Networks for Anomaly Detection in Crowded Scenes. IEEE Transactions on Information Forensics and Security.
4. Mohammadi, S., Perina, A., Kiani, H., and M. Vittorio, 2016. Angry crowds: Detecting violent events in videos. In ECCV.
5. Kamijo, S., Matsushita, Y., Ikeuchi, K.and M. Sakauchi. Traffic monitoring and accident detection at intersections. IEEE Transactions on Intelligent Transportation Systems, 1(2):108–118, 2000.
6. Sultani, W., and Choi, J. Y., 2010. Abnormal traffic detection using intelligent driver model. In ICPR.
7. Lu, C., Shi, J., and Jia, J., 2013. Abnormal event detection at 150 fps in matlab. In ICCV.
8. Zhao, B., Fei-Fei, L., and Xing, E. P, 2011. Online detection of unusual events in videos via dynamic sparse coding. In CVPR.
9. Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In CVPR, 2015.
10. Mohammad Saberian Zhaowei Cai and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In ICCV, 2015.
11. Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In CVPR, pages 152–159, 2014.
12. Xiaogang Wang, Meng Wang, and Wei Li. Scene-specific pedestrian detection for static video surveillance. TPAMI, 36(2):361–374, 2014.
13. Ross Girshick. Fast r-cnn. In ICCV, pages 1440–1448, 2015.
14. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, pages 580–587, 2014.
15. Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multiscale orderless pooling of deep convolutional activation features. In ECCV, pages 392–407. 2014.
16. Li, J., Liang, X., Shen, S., Xu, T., Feng, J. and Yan, S., 2018. Scale-aware fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia, 20(4), pp.985-996.
17. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), p.436.
18. Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.
19. Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).
20. Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).
21. Mahadevan, V., Li, W., Bhalodia, V. and Vasconcelos, N., 2010, June. Anomaly detection in crowded scenes. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 1975-1981). IEEE.
22. Kim, J., and K. Grauman., 2009. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In CVPR, pages 2921–2928.
23. Mehran, R., Oyama, A., and Shah, M., 2009. Abnormal crowd behavior detection using social force model. In CVPR, pp. 935–942.
24. He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).
25. UCSD Anomaly Detection Dataset, http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm