International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April, 2020

Video Object Detection through Traditional and Deep Learning Methods

Sita M. Yadav, Sandeep M. Chaware

Abstract*: Object detection in videos is gaining more attention recently as it is related to video analytics and facilitates image understanding and applicable to . The video object detection methods can be divided into traditional and deep learning based methods. Trajectory classification, low rank sparse matrix, background subtraction and object tracking are considered as traditional object detection methods as they primary focus is informative feature collection, region selection and classification. The deep learning methods are more popular now days as they facilitate high-level features and problem solving in object detection algorithms. We have discussed various object detection methods and challenges in this paper.

***Keywords* : Video Object Detection, Deep Learning Methods**

I. INTRODUCTION

Computer vision is a field in which, object detection from the video sequences is an interest point for many vision based application like, video surveillance, traffic controlling, action recognition, driverless cars and robotics. The task of object detection includes localization and classification. From video frames data is extracted to predict the objects in which task of drawing a bounding box around one or more object is called localization and task of assigning label is classification. The object detection from video sequences can be based on feature, template, classifier and motion. Various papers have discussed about role of moving camera and fixed camera in object detection. But object detection in videos which capture using moving cameras is less and work is still going on. Object detection becomes primary requirement for computer vision which helps in understanding semantic of images and videos.

II. LITERATURE SURVEY

In [1] the author introduced method based on single deep neural network for detecting objects. The approach is based on SSD which use aspect ratio and scales for feature map, performance can be improved by using RNN. In [2], the authors have proposed a Region Proposal Network (RPN) which work on detection network with full-image convolutional features, hence gave cost-free region proposals. This paper showcases a deep learning based object detection method which achieves speed of 5-17 fps. [3] have proposed a framework by using object detection, classification and semantic event description. The event is analyzed by integrating the object detection and scene categorization. The system can be improved by automatic scene learning methodologies.

The authors of [4] have proposed methods and architectures to understand videos. The architecture is given for automatically categorization and caption in the video. The system implemented on temporal feature pooling (TFP), 3D Convolution, frame majority and LSTM for classification. Microsoft multimedia dataset used, the output is the predicted video categories and video captioning. Better dataset cleaning is required along with focus regions. One frame per second extracted from video which may probably missed some important information. The various detection algorithms are explained using given algorithm but accuracy of detection is not discussed. [5] proposed a system to detect moving objects using background subtraction, edge detection and geometrical shape identification**.** If the object is moving in speed then this system does not give accurate result. [7] Suggested pedestrian detection method which separates the foreground object from the background by utilizing image pixel intensities. The foreground edges are enhanced by high boost filter. [8] the authors put forward object detection system using CNN. The neural network algorithms are able to handle the occlusions and camera shake problems, with use of frame difference method. However, proper analysis of training model is required. [9] introduces BMA (Block matching algorithm) for moving object detection. This method divide the video frames into non-overlapping blocks then matching is done in reference frame. The computational time for BMA is low and robust. However, further study is required for lossless compressed video based Background Subtraction (LIBS) method is used. [14][15] have given state of art region based object detection methods.

III. FACTORS AFFECTING OBJECT DETECTOR

The object detection requires to identify the features that impact performance of detector with framework. Based on literature survey the various factors which affect detector performance are feature extractor, threshold decision for loss calculation, boundary box encoding, training dataset, data augmentation, localization factors and feature mapping layers.

Revised Manuscript Received on April 25, 2020.
* Correspondence Author
Sita M Yadav*, Computer Department, AIT Pune, Research Scholar at PCCoE, Pune, Maharashtra, India. Email:yadav.sita1@gmail.com
Dr. Sandeep M. Chaware, Computer Engineering Department, MMCOE, Pune, Research Guide at PCCoE, Pune, Maharashtra, India.
Email: sandeepchaware@mmcoe.edu.in

IV. VARIOUS OBJECT DETECTION METHODS

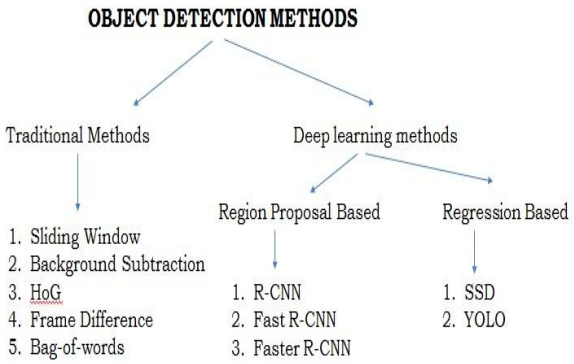


Fig 1: Various object detection methods

A.TRADITIONAL METHODS

- **Sliding Window Model** (Region Selection)[10]
  The sliding window method consider fix sized window to identify the objects which are present at different places in an image. The windows are scan left to right, top to bottom etc. for finding potential objects. The candidate windows give large number of redundant windows.
- **HOG, SIFT, Haar**[14]

The Hough transform identify imperfect objects within a class by voting procedure. This voting based on parameter space is carried out, which gives object candidates with accumulator space constructed explicitly for computing the Hough transform. HOG transform can detect shape and category. However, due to lot of diversity in object appearances the robust feature extractor is a challenge.

- **Frame Difference Method**
  This is one of the widely used method which work on the concept of pixel difference.
- **Background Subtraction[6]**
  These is one of the most popular method used for object detection.

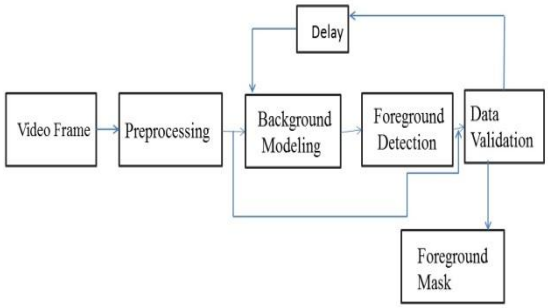


Fig 2. Background subtraction method[6]

The model statistically extracts the features from background, then some features are extracted from current frame and then corresponding background features are found. Thus, this method has background modeling, object detection and background updating.

- **Classification**

The classification process needs a strong classifier give target object representation in hierarchical, semantic and informative formats. Supported Vector Machine (SVM) [12] and DPM [13] are popular classifiers.

In general , object detection methods are divided into two categories, frame difference model, background subtraction and Hough transform method which adopts feature extraction with mathematical model , and second sliding window , deformable part model , feature extraction with hand engineered classifier feature to detect object. The above two categories are not efficient as they cannot bridged the gap by combining manually engineered features and discriminatively trained shallow models. They are redundant, inefficient and in accurate.

B.OBJECT DETECTION USING NEURAL NETWORK

The emergence of neural network algorithms better precision in object detection methods is achieved. Object region proposal, feature extraction and classification are major steps involved in object detection. These models are further divided into region proposal and regression based models.

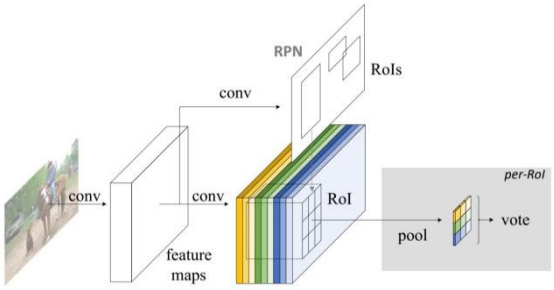


Fig 3. RPN, RoI in CNN[2]

A. MODEL BASED ON REGION PROPOSAL

- **R-CNN[14]**

R-CNN is a region-based **C**onvolutional **N**eural **N**etwork proposed by Girshick [14] in 2014. These methods include region segmentation method with region proposal to identify objects. The method apply selective search algorithm and get 2000 region proposals, then apply CNN on region proposals after that SVM is used for classification. R-CNN give 58.5% accuracy in detection but it is slow due to selective search. Hence R-CNNs are not widely used in actual applications.

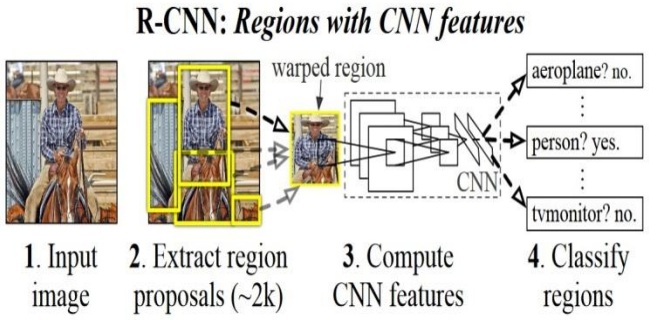


Fig 4. R-CNN[14]

- **Fast R-CNN [15]**

The Fast R-CNN avoids region proposal like R-CNN. The convolutional feature map, help in identify the region of proposals by using selective search algorithm, then ROI pooling is used to map the feature from proposed region, the ROI pooling gives fix size vectors which is essential for fully connected CNN. SVM is replaced by softmax layer. Suppose the selective search generates *n region* proposal. The CNN give different region of interest as different shapes are present in image. The Fast R-CNN uses CNN output for RoI pooling , the output shape is determined by using fully connected layer. The softmax regression is used for category prediction.

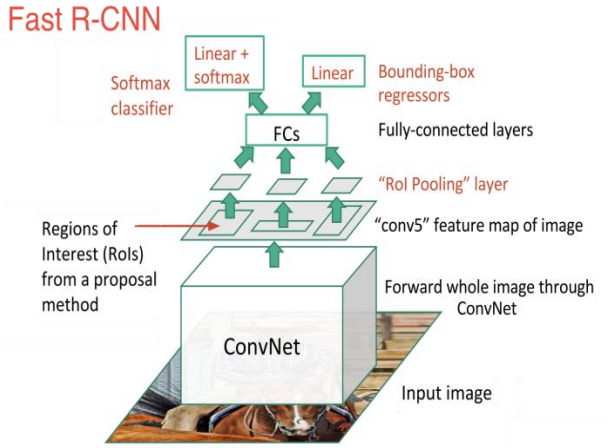


Fig 5. Fast R-CNN[15]

- **Faster R-CNN [2]**

[2] proposed modified version of Fast R-CNN named Faster R-CNN , which uses end to end connected framework. Faster R-CNN uses region proposal network (RPN). The RPN divide the feature layer of an initial CNN into regions and give "objectness" score for that region.

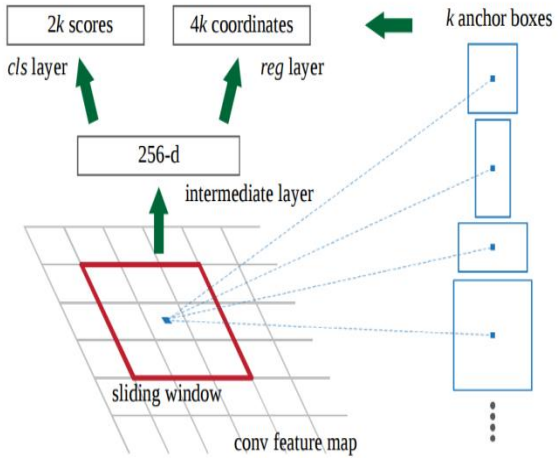


Fig 6. Feature Map[2]

The RPN output gives bounding box coordinate it does not classify any object. If the threshold of anchor box is more than that box coordinates are forwarded for region proposal. After getting region proposals, the softmax layer is used for classification.

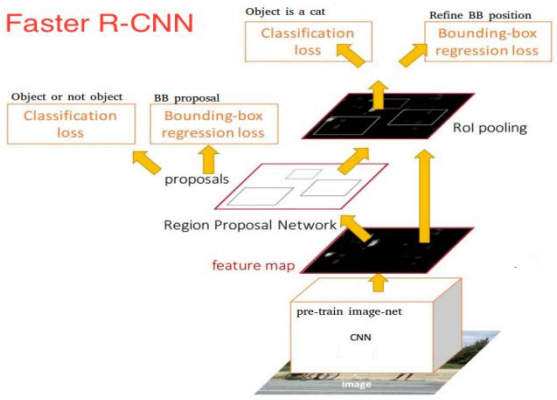


Fig 7. Faster R-CNN[2]

Experimentally it is proved that the performance of faster R-CNN is better than Fast R-CNN. Faster R-CNN proposes 300 regions as it use RPN. The mAP of Faster R-CNN has been raised to 70.4% in PASCAL VOC2012 and 42.7% in MS COCO as compared to Fast R-CNN.

- **R-FCN [16]**

Dai proposed the R-FCN, which adapts the concept of "Increase speed by maximizing shared computation". The R-FCN uses generate candidate region of interest by using region proposal network. The R-FCN is fully connected hence it share 100% of the computations among the all convolutional layers. The positive sensitive score method is used in R-FCN to give position of object class.

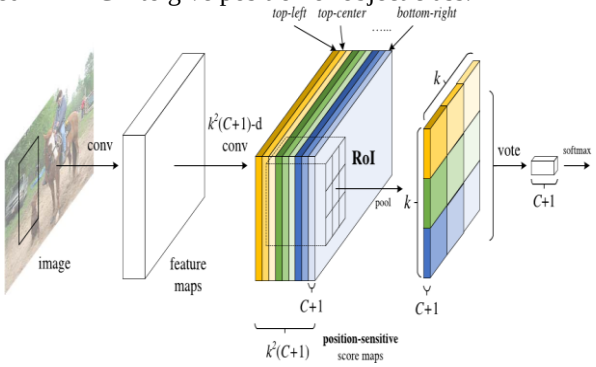


Fig 8. R-FCN[16]

The mAP of R-FCN in PASCALVOC2012v and MS COCO dataset is 77 % and 49.2% .

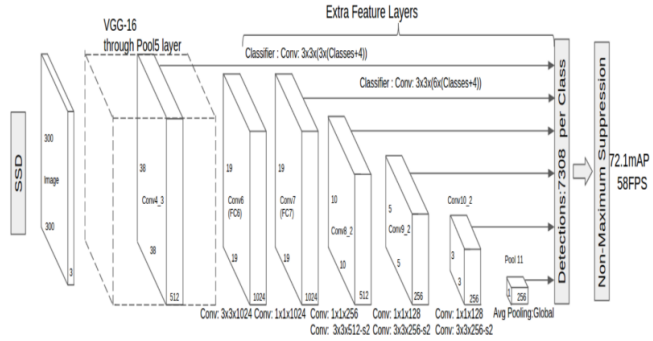
B. Model Based on Regression (Regression based framework)

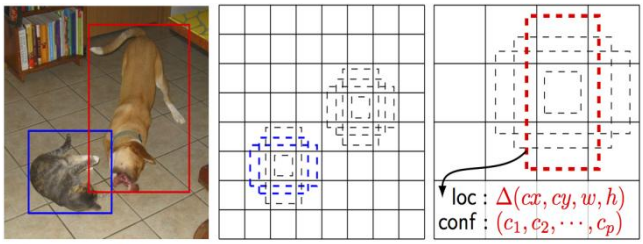
The region proposal based frameworks include region proposal algorithm , feature extraction with CNN, classification and bounding box regression which are separate stages , hence processing time for each stage is different which affects real time performance. Hence one step solutions are required the two famous approaches for regression framework are SSD and YOLO.

- **SSD [1]**

SSD (Single shot multibox detector) proposed by Liu Wei. In earlier methods the model is implemented by performing region proposals and classifications. SSD perform above steps in a "single shot," also it predicts the bounding box and the class of object. SSD group

overlapping boxes by using non-maximum suppression method. It keeps the boxes with highest confidence and discards others.





(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Fig 9 . Region proposal SSD[1]

The object detection performance accuracy of SSD is 74.9% on PASCAL VOC 2012 and this model satisfy the real time object detection requirement.

- **You only look once [17]**

You only look once was given by Redmon et.al. work for real time object detection and by using end to end training. The feature map is used to predict confidence for multiple categories and bounding boxes. The confidence scores can be calculated by Pr(Object)*IOU_{pred}^{truth} , which indicate how likely there exist objects (Pr(Object) ≥ 0) and confidence of this prediction is IOU_{pred}^{truth} . At the same time class probabilities for each grid cell should be predicted, Pr(Object)*IOU_{pred}^{truth} * Pr(Class | Object) = Pr(Class) * IOU_{pred}^{truth}

YOLO, YOLOV2 and YOLO V3 has proven efficiency in real time object detection. The accuracy of YOLO on PASCAL VOC 2012 is 73% and with MS COCO it is 33%.

Table 1 : Result comparison of R-CNN[14],Fast R-CNN[15], Faster R-CNN[2], YOLO[17] and SSD [1], dataset PASCALVOC 2007

Methods	Dataset	mAP	FPS	Real time speed
R-CNN (Alex)[14]	PASCAL VOC 2007	58.4	-	No
R-CNN(VGG)[14]	PASCAL VOC 2007	66.0	-	No
Fast R-CNN[15]	PASCAL VOC 2007+2012	68.1	0.5	No
Faster R-CNN (VGG)[2]	PASCAL VOC 2007	69.9	7	No
YOLO [17]	PASCAL VOC 2007	63.7	45	Yes
FAST YOLO [17]	PASCAL VOC 2007	78.6	155	Yes
SSD300 [1]	PASCAL VOC 2007	77.6	56	Yes

V. CHALLENGES IN VIDEO OBJECT DETECTION

The major objective of moving object detection is to detect moving objects present in video frames which are extracted from videos captured through fixed/moving cameras. The task has many difficulties and challenges as given below,

A. Moving Object Annotation

The spatio-temporal relationship of pixels is required for object detection. The background identification is a tedious task as it include forest, forest fire, hurricanes, water etc. The objects like hand, fingers which have complex shape cannot be identified by simple geometric shape representation.

B. Illumination Variation

The change in illuminations can be seen due to the change in source of light, reflections, outdoor conditions or any disturbance in light source.

C. Change in appearance of moving object

The change in appearance of object is a major concern as objects move in 3D spaces but projection of 3D to 2D may cause change in appearance.

D. Occlusion

The occlusion may appear due to object overlapping.

E. Complex Background

The background in object detection should be understood properly as changes in background affects object detection.

F. Shadow

The shadow creates complications in object detection. Various studies [16] proved that the methods based on Gaussian mixture model, HSV, better segmentation methods for moving objects can handle shadow in object detection properly.

I. Real Time Performance Aspect

Many of earlier methods were not able to give performance in real time. The deep learning methods detect object in real time, but in deep learning large amount of computations are involved which make this task tedious and affects real time object detection efficiency.

J. Robustness

The robustness is more in deep learning algorithms as compared to traditional methods. Train the deep learning model in such a way that improves the detection abilities with robustness. SSD has improved ability to identify the small objects, YOLO is faster mechanism for identifying object in videos.

VI. RESULTS

The comparison of various deep learning based object detection methods is given below indicating YOLO and SSD models give better results for real time object detection.

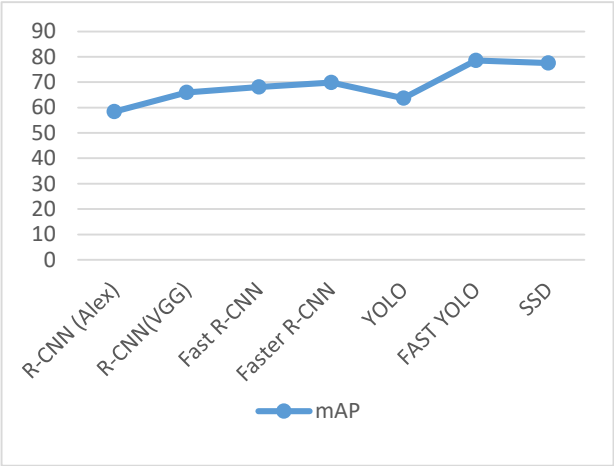


Fig 10: Accuracy Comparison [1][2][14][15][16][17]

VI. FUTURE DIRECTIONS

The traditional and deep learning based approaches detect objects with certain level of accuracy but still there is lot of scope for future work.

A. To identify small objects accurately it is required to localization of small objects.
B. Correlation of different tasks within and outside object detection should be done to achieve multi model information fusion and optimization.
C. Scale invariant detector creation for various robust scale adaption
D. Use of cascade networks
E. Adaption of 2D object detection method by 3D object detection.

VII. CONCLUSION

In video object detection the deep learning algorithms are popular nowadays as they have capability of dealing with occlusion, scale transformation and background changes. This paper discussed about the details of various traditional and deep learning based object detection methods with their limitations. Further the challenges in video object detection are discussed. The video object detection need more robust methods to perform in real time scenario hence more research is needed in use of deep learning methods along with upgraded computer hardware to achieve the existing challenges of real time object detection.

REFERENCES

1. W. Liu, D. Anguelov, D. Erhan, C. Szegedy et. al., "SSD : Single Shot MultiBox Detector", ArXiv: 1512:02325, Dec 2016.
2. S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6) , June 2015.
3. J. Sun, J. Wang, T. Yeh,"Video Understanding: From Video Classification to Captioning", Stanford university,2016.
4. Liu, C. Hu, "Video event description in scene context", Journal of neuro computing, Elsevier, 2013.
5. H. Belhani, L.Guezouli, "Automatic detection of moving objects in video surveillance", 2016 Global Summit on Computer & Information Technology, 2016.
6. Soundrapandiyan, R. ,Mouli, P.V.S.S.R.C., "Adaptive Pedestrian Detection in Infrared Images Using Background Subtraction and Local Thresholding". Procedia Computer Sci.58, 2015.
7. M. Chihaoui, A. Elkefi, W. Bellil, " Detection and Tracknig of the moving objects in a video sequence by geodesic active contour." 13th International conference on Computer Graphics, Imaging and Visualization , IEEE, 2016.
8. B. Tian, L. Li， Y. Qu， Li Yan, "Video Object Detection for Tractability with Deep Learning Method", Fifth International Conference on Advanced Cloud and Big Data 2017.
9. S. Safie, A A Samah, G. Sulong , "Block Matching Algorithm for Moving Object Detection in Video Forensic", 2017 6th ICT International Student Project Conference , DOI: 10.1109/ICT-ISPC.2017.8075330 , 2017.
10. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.
11. M. Ulrich, C. Steger, A. Baumgartne, "Real-time object recognition using a modified generalized Hough transform," Pattern Recognition, vol. 36, 2003.
12. Cortes and V. Vapnik, "Support vector machine," Mach. Learn., vol. 20, no. 3, 1995.
13. P. F. Felzenszwalb et al., "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, Sep. 2010.
14. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. CVPR, 2014.
15. R. Girshick, "Fast R-CNN," in Proc. ICCV, 2015.
16. Y. Li, K. He, J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," Advances in Neural Information Processing Systems, 2016.
17. J. Redmon, S. Divvala, R. Girshick, et al, "You only look once: Unified, real-time object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

AUTHORS PROFILE

Sita Yadav is currently working as Assistant Professor in Computer Engineering Department at Army Institute of Technology affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India. She is currently pursuing Ph.D. in Computer Engineering from PCCOE-SPPU. Her domain of interest is Computer Vision, Image Processing, Machine Learning and Artificial Intelligence.

Dr. Sandeep M. Chaware is currently Professor in Computer Engineering Department at Marathwada Mitramandal College of Engineering affiliated to Savitribai Phule Pune University, Pune, and Maharashtra, India. He has many paper publications in International/ National Conferences and Journals. His domain of interest is Natural Language Processing, Multilingual Database Processing, Mobile Computing, Image Processing, Machine Learning and Artificial Intelligence. He is member of ISTE, CSI and Advisory Board Member for IRJMRS . He has served as Technical Program Committee member and Reviewer for Several International Conferences and Journals.