Deepanshu Wadhwa et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 1, March 2018, pp. 173-177

Technical Survey on Object Detection Methodologies

Deepanshu Wadhwa[1*], Devina Shah[2*],Utkarsh Maharana[3*],Tanvi Dhingra[4]
Department of Information Technology, Northern India Engineering College, New Delhi
[1]deepanshuwadhwa97@gmail.com
[2]devina.shah@yahoo.com
[3]utkarshm27@gmail.com
[4]tanvi.dhingra@niecdelhi.ac.in

Abstract **— Object detection is a significant and an elementary problem in the area of computer vision implementation due to the problem of tracking of entities which can rise due to several internal and external reasons like distortion, motion of camera, motion blurring and occlusion. This paper provides a literature survey on different state-of-the-art object detection and tracking strategies so as to lead to a decrease in the tracking divergence.**

Keywords **— Computer vision, object detection and tracking, image processing, CNN (Convolution Neural Network), R-CNN, faster R-CNN, ReLU (Rectified Linear Unit), YOLO (You Only Look Once), background subtraction, SAD (Sum of Absolute Difference), DNN (Deep Neural Network), SSD (Single Shot Multibox Detector).**

I. INTRODUCTION

In the computer perception applications, object detection is a foundational area which requires the analysis, distinguishing and recognition of an object with the help of computers [2]. Object detection has several benefits which come with certain hindrances such as the problem of reducing the deformation present in any sequence and motion tracking, which causes numerous issues in practical areas and produces various problems leading to the outcome of noise as well as disturbance. [2] To remove all these disadvantages, many entity-tracking applications are developed. In this paper, we analyse some important algorithms for object detection. The pre-existing techniques that have been developed for this particular area include techniques like Histogram of Oriented Gradients (HOG) [1], Local Binary Pattern (LBP) [16] technique, Co-tracking through Support Vector Machine (SVM) [15] to name a few.

II. SURVEY

A. *Intelligent Service Robot involving Deep Learning*

The algorithm in this research [3] is used to make a service robot more capable of object detection and recognition in complex scenes, aiming at improving the accuracy and real-time performance. The algorithm is a deep learning based, end to end object detection algorithm. It is an integration of the local multi branch deep convolutional neural network [4], which enhances the feature representation proficiency of the model by improving the feature enhancing module function and combining it with the anchor point mechanism, the object class and position regression prediction, which is carried out on the multi-layer feature map. This is followed by the realization of natural multi-scale detection and recognition on multiple receptive fields, after the local and global features are fully fused [18]. Everything is compiled and executed on a network acceleration module for GPU parallel. All the convolutions in the local multi-branched modules as well as the reduction layers use rectified linear activation (ReLU). The network is 22 layers deep, only including the layers with parameters, and is designed to be run on individual devices including even those with computational resources, specifically with low memory footprints. The end to end regression structure comprises of :

a)R-CNN algorithm [5]- It combines the region proposals with CNN for the first time.

b)Faster R-CNN algorithm [6]- It joins region proposal and CNN classification together, using end to end neural networks for target detection.

c)YOLO algorithm [7]- To overcome the limitations of the Faster RCNN method, i.e. the speed is not close enough to the real-time effect, YOLO algorithm [7] was proposed, which increases the speed of the algorithm greatly, but compromises the accuracy of the algorithm. This algorithm extracts the feature of the input image by deep CNN, and distributes the mes7h into the last layer feature map of the final network, then directly carries on the boundary and the class regression on the real picture. But it comes with its own defects, i.e. each grid only predicts only one object which makes the scalability of the object relatively sensitive.

The algorithm proposed in this paper tries to overcome the shortcomings of the YOLO algorithm in terms of accuracy. This algorithm uses local multi-branch deep convolutional neural network as its basic network structure and the general structure as the basic network. The other layers are added on the basis of these two layers. This increases the depth of the network, increasing the accuracy to detect more objects at different scale and ratio. For

each point on the feature map obtained by the convolution operation, B boxes are generated of different scales and ratio, similar to anchor in Faster R-CNN. For example, a layer of dimension 'axb' with t channels, the basic element for predicting parameters of a potential detection is 3x3xt small kernel that produces either a score for a category or a shape offset relative to the box coordinates. And each 'axb' coordinate where the kernel is used, the output is produced. Let y^t_{mn}= {1,0} be a parameter for matching the m-th box to the n-th ground truth box of category t. In this matching strategy, we can have $\sum$ ytmn≥ 1. The overall objective loss function is a weighted sum of the localization loss(los) and confidence loss(con):

$O(y,d,p,i) = 1N(O_{con}(y,d) + \mu O_{los}(y,p,i))$ (1)

Here, N – number of matched boxes; if N=0 then loss = 0; Smooth O1 – localization loss between the predicted box (p) and the ground truth box (i) parameters.

B. Fuzzy Based Technique [14]

Background subtraction is one of the active areas in the field of computer vision. It is the process of separating background from the captured frame. Background can be either static or dynamic. Various factors affect the subtraction process such as illumination, occlusion, noise etc. It is most difficult in case of videos, which constitutes of sequence of frames which move at a very high frequency. But, the contents of the two consecutive frames are very similar. Object detection is carried out via the background subtraction algorithm, in which we set up an initial background with the help of background modelling and then subtracting the current frame from a previous frame to detect the objects in motion. Due to external factors such as illumination, shadows etc, a better improvised alternate background subtraction technique is required. The basic foreground detection is given by [8]:

$BF^m(a,b) = \{1; \quad PF^m(a,b) - BF^{m-1}(a,b) | \geq mh_s$ (2)

$\{0; \quad$ otherwise

Where PF^m – present frame at time m

BF^{m-1} – Background frame at time m-1

mh_s – is the predefined threshold

$BF^m(a,b)$ – result of background subtraction

C. A Block-Based Background Model

The general aim of the research paper [9] is to improve the accuracy with which the moving objects can be detected in a video sequence when the object stays in the sequence for long, also countering the problem of illumination change, which has been a known factor affecting the background subtraction method, that is a method used for generating a background model. This approach has been proven to be effective experimentally and has low computational cost, which makes it feasible for embedded systems as well. The major task in detection of an object is to generate the background model that can effectively be distinguished from the moving object in the frame. The paper [9] runs different tests and processes to make sure the background modelling is done right.

There are two approaches of background modelling – a pixel based approach and a block based approach. The paper is based on background subtraction method in combination with block based background generation method, where the initialization or the formation of the first background is done by determining the SAD values between two consecutive images. This first background is then evaluated and updated on the advent of new frames, by determining the entropy of each block. These two techniques are used for generation of an accurate background frame. SAD is a measure of the similarity between image blocks. It is determined by calculating the absolute difference between each pixel in the original block and the corresponding pixel in the block being used for comparison.

To minimize the information in each frame and to get rid of the noise to the maximum extent along with the false modelled pixels, we compute the Structure-Texture Decomposition [17] of different absolute images. The structure component is then used for further processing in detection of the moving object. Talking about the proposed method, background initialization involves the following steps. The first step of the foreground detection method, is to generate a background. A background is basically the set of pixels that remain stable in a video sequence. To achieve a background model, an initial background image in generated first.

To do so, the images of the sequence are divided into blocks (NxN). The SAD between two consecutive images of the sequence is calculated and the blocks containing the minimum SAD value, calculated from 100 initial images, is then chosen to create an initial background. This process is followed by the entropy evaluation and elimination of noise which is done before the starting of thresholding procedure. For detecting the moving object, we apply the following formula on the structure component and subsequently form the binary image H(x,y). (3)

$$SAD_s^{(p,q)} = \sum_{x=1}^{N}\sum_{y=1}^{N} |I_{S_t}^{(p,q)}(x,y) - B_{S_t}^{(p,q)}(x,y)|$$

I_{St} – Structure component of current frame

B_{St}– Structure component of background frame

(x,y) - Pixel coordinates

$$H_t^{(p,q)} = \begin{cases} 1, & if\ SAD^{(p,q)} > Q \\ 0, & otherwise \end{cases} \quad (4)$$

where, Q = Mean of SAD values of blocks of foreground detection image.

D. A SSD-Based Object Detection System

The paper [10] focuses on the application of deep learning using a distributed system combining the cloud and mobile devices for detection and recognition of road vehicles. The distributed system consisting of the cloud and mobile devices is implemented using the scheduling algorithm. The paper [10] attempts to tackle the issue of deep learning-based video processing applications quickly draining the resources of mobile devices like energy and memory. However, if as a solution to this, all the DNN models are placed in the cloud, the transmission of data will get interrupted due to network issues and the transmission speed too is difficult to guarantee.

Thus, the paper [10] puts forward a terminal to cloud-based system that selects the model in the cloud or the one on the mobile devices adaptively. KITTI [11] dataset was used for training the model and SSD [12] model variants were generated in Caffe [13]. On the side of mobile devices, the Surface Texture class was used in Android (3.0 or later) to process the video stream from the camera.

In the paper [10], a new dataset was used to retrain and improve the training methods of the model proposed by Liu, which combines YOLO [7] framework's regression theory with Faster R-CNN's [6] anchor mechanism to maintain the fast speed of YOLO [7] and also to ensure the same accuracy of prediction as Faster R-CNN [6]. The system design [10] is as follows:

The entire system consists of two main parts: the Linux based cloud server and an Android based smartphone. There are five primary modules in the system:

• Video stream and acquisition module to collect video data of Android camera
• Scheduling module for determining the operating position of models (smart phones or cloud)
• Video data transmission module for transmission of data between the Android and the Linux side
• Video frame detection and recognition module which runs on Android or Linux side to process video images for each frame for object detection and recognition
• Video display module for playing the processed video on the Android side.

The program starts once the user clicks the application icon on the mobile. The application opens the camera of the device and reads the video data being captured. The model dispatch interface of the program then selects the most appropriate model variant by detecting the current power, memory and speed of the device. For this, the scheduling algorithm is used which determines which model (cloud or the mobile) to use based on the following formula where the variables are obtained by making the server continuously send the request data:

$$x = \frac{Y_{i+1} - Y_i}{t} \quad (5)$$

where, x is the average speed per second; Yi is the current total amount of data received; Yi+1 is the total amount of data received, calculated once every other time step; t is the interval time step. If the size of video data per frame is assumed to be S, and the one-way transmission delay is D, when $x<S/D$, the network is judged to be poor and the processing is done on the device itself. LRU (Least Recently Used) algorithm is used to minimize the number of times a model is load into or evicted from the memory so as to reduce the power and time consumption.

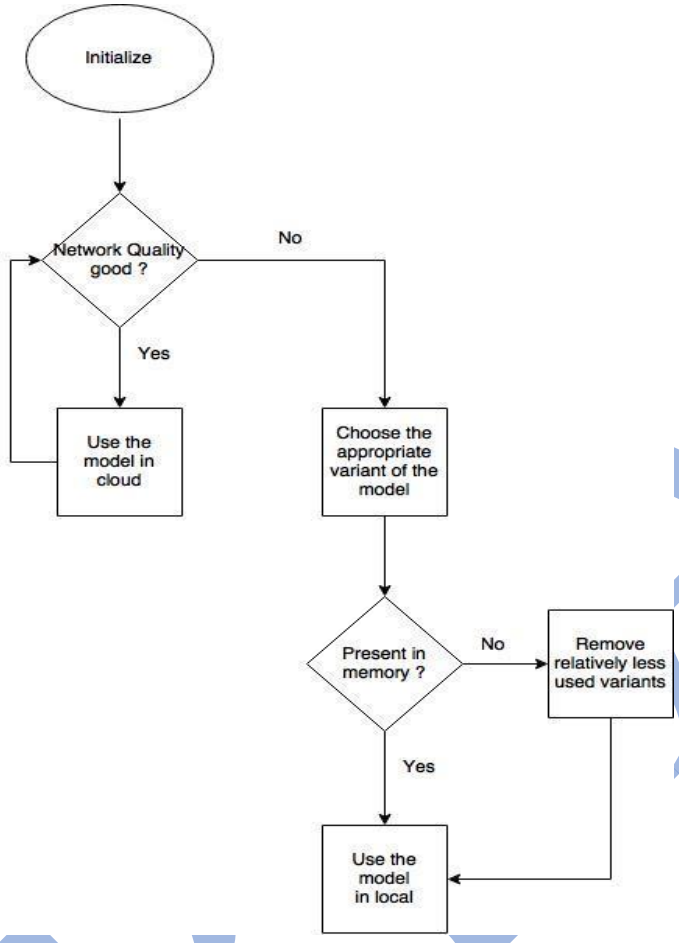

Fig. 1 scheduling algorithm working

DNN based object detection model SSD [12] is used for object detection and KITTI [11] dataset is used for training. SSD [12] is based on a forward propagation CNN, which results in a series of fixed size bounding boxes with each box containing the possibility of an instance of the object after which non-maximum suppression is used to obtain final predictions.

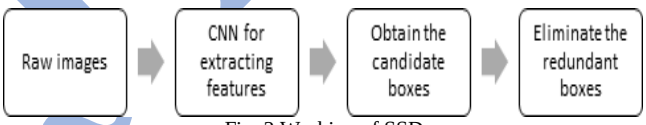

Fig. 2 Working of SSD

III. Conclusion

In this paper many object detection and tracing algorithms were dealt with, and several of their shortcomings and drawbacks were mentioned in each and every methodology. The most commonly observed pattern was the trade-off between speed and accuracy across all the algorithms studied. Each algorithm made use of a different approach to tackle the objective of object detection with each approach carrying along several advantages and drawbacks associated with it. In future scope, advanced study will be carried out to find an efficient strategy to lower the tracking drift and to study the possibility of further improvement in the existing models.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", In CVPR, 2005.
[2] K. Rasool Reddy, K. Hari Priya, N. Neelima, "Object Detection and Tracking -- A Survey"2015 International Conference on Computational Intelligence and Communication Networks (CICN)pp. 418 – 421, 2015.
[3] Yanan Zhang, Hongyu Wang, Fang Xu, "Object detection and recognition of intelligent service robot based on deep learning" 2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM) pp. 171 – 176, 2017
[4] Krizhevsky A, Sutskever I, Hinton G. E., "ImageNet classification with deep convolutional neural networks" Advances in neural information processing systems, pp.1097-1105, 2012

[5] Girshick R, Donahue J, Darrell T, et al. Rich features hierarchies for accurate object detection and semantic segmentation, IEEE conference of computer vision and pattern recognition., pp. 580-587, 2014
[6] Ren. S, He. K, Girshick. R, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks", Advances in neural information processing systems, pp. 91-99, 2015.
[7] Redmon J, Divvala S, Girshick R, et al. "You only look once: Unified, real-time object detection." IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.
[8] T. Bouwmans, "Background subtraction for visual surveillance: A fuzzy approach," Handbook on soft computing for video surveillance, vol. 5, pp. 103–138, 2012.
[9] Omar Elharrouss, Abdelghafour Abbad, Driss Moujahid, Hamid Tairi, "Moving object detection zone using a block-based background model, Volume: 12", Issue: 1 IET Computer, pp. 86 – 94, 2018.
[10] Buren Qi, Mengfei Wu, Lin Zhang, "A DNN-based object detection system on mobile cloud computing", 2017 17th International Symposium on Communications and Information Technologies (ISCIT), pp. 1 – 6, 2017.
[11] Geiger A, Lenz P, Stiller C, et al, "Vision meets robotics: The KITTI dataset," The International Journal of Robotics Research, 32(11), pp. 1231-1237, 2013.
[12] Liu, Wei, et al, "SSD: Single shot multibox detector," European Conference on Computer Vision, Springer International Publishing, pp. 21-37, 2016.
[13] K Jia, Yangqing, et al, "Caffe: Convolutional architecture for fast feature embedding," Proceedings of the 22nd ACM international conference on Multimedia, ACM, pp. 675-678, 2014.
[14] Bibhu Prasad Das, Priyanka Jenamani, Subrata Kumar Mohanty, Suvendu Rup, "On the development of moving object detection from traditional to fuzzy based techniques" 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 658 – 661, 2017
[15] Hui Zheng, Jun-xia Zhang, "Study on fault diagnosis of SVM for mechanical and electrical product based on improved conjugate transformation," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 1888 – 1892, 2017.
[16] Meghana Dinesh Kumar, Morteza Babaie, Shujin Zhu, Shivam Kalra, H. R. Tizhoosh," A comparative study of CNN, BoVW and LBP for classification of histopathological images", 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1 - 7, 2017.
[17] Vese, L.A., Osher, S.J, "Modelling textures with total variation minimization and oscillating patterns in image processing," J. Sci. Comput., 19, (1–3), pp. 553–572, 2003.
[18] Gaurav Kumar, Pradeep Kumar Bhatia, "A Detailed Review of Feature Extraction in Image Processing Systems", IEEE 4th International Conference on Advanced Computing & Communication Technologies, pp. 5-12, Feb. 2014.

