ISSN: 0974-5823 Vol. 7 No. 5 May, 2022

International Journal of Mechanical Engineering

Multi-target Detection and Tracking in CCTV using Deep Learning Techniques

Rohini Chavan

E&TC Department

Vishwakarma Institute of Information Technology

Sanket Patil

E&TC Department

Vishwakarma Institute of Information Technology

ABSTRACT

Video object and human action detection are getting used today in many fields, such as video surveillance, face recognition and many more. Video object detection focuses on object classification, while human action detection is recognition of activities done by humans.In this paper we are focusing on video detection methods using supervised learning which is a part of deep learning itself. Here heterogenous training data and data augmentation is used to improve detection in CCTV surveillance. Objects' spatial transformation parameters are proposed to be used to predict the evolution of camera parameters and based on that tune the detector for better accuracy. For target detectionFaster R-CNN algorithm in deep learning is being used here. The research in this paper helps for multi target detection in CCTV footage. We have use dataset downloaded from internet for validation of algorithm. For evaluating R-CNN algorithm performance various matrices are being used.

Index Terms– CCTV, R-CNN, motion blur, spatial

I. INTRODUCTION

Object detection has much more importance in research community because of its practicality. Not only this but also it has use in smart video surveillance solutions also. This can be used in traffic controlling on roads, crowd controlling in busy areas such as railway stations, airports, etc. Object tracking method is not an easy task that too from CCTV cameras. Challenges such as motion blur, video defocus, part occlusion, etc. are faced. We all have seen that cameras generally changes focus from one object to another based on the movement and distance. Not only this but also the shape of theobjects gets changed with the change in distance of that particular object with respect to CCTV camera.

Most of the existing detection methods were mainly focused on building a robust object appearance model, working on feature representation and classifier construction. However, most of these classifiers are limited by their shallow structures while object appearance variations are complex and time-varying[1]. Recent advancements in deep learning have led to a new generation of object detection and localization methodologies that outperform the traditional methods. They rely on automatically learning discriminative features via a multi-layer convolutional neural network, thus, alleviating the need for handcrafted features[2].

The existing video detection methods are operated on frames. Most of the existing video detection methods are to decompose the video into frames, and then use the image detection method to detect the frames. Therefore, the speed of video detection depends on speed of image detection[3]. While some methods directly operate on the video, however, these methods are also frame-based. They operate adjacent frames by using specific algorithms. Therefore, for video detection, image detection methods are still important.

Histogram of Oriented Gradients, Scale-Invariant Feature Transform, Haar-like feature, etc. Haar-like feature is from Haar wavelet, which is a kind of square-shaped function. These methods were used to extract features of the images, and later they were used for detection[4].

In the deep learning tasks, some loss functions can also be regarded as the classifiers, since only the class which conforms to the loss function can be detected and recognized. These loss functions include cross-entropy function and some loss functions customized by researchers themselves.

Before deep learning, local feature extraction methods such as SIFT, HOG, etc. did not have the ability of feature translation invariance. The possible reason may be that the features extracted by these methods may be simpler than the deep learning methods[5].

In this work, a multiple-object detection framework for tracking by detection applications that confronts the challenges of real-world CCTV videos is proposed. It is based on a state-of-the-art detection and localization object framework that is trained offline to facilitate a tracking-by-detection paradigm[6].Generally, video detection is much more complicated process than image

detection as there are multiple frames in videos. Let's take an example of CCTV footage, there are many challenges such as low camera quality and other environmental factors. Specially the moving background is the real challenge.

II. LITERATURE SURVEY

CCTV videos does contain severely blurred objects because of low video quality and fast PTZ operations. Especially the motion blur is a major challenge for object detection in CCTV footages. In a single frame, motion blur gets translated to degraded appearance information and reduced ability to accurately localize the position of an given object. While de-blurring the methodologies show good results, they have a high computational cost, and they further degrade their appearance. Deep learning systems have been recently shown to achieve impressive performance in benchmark datasets for object detection. However, in challenging CCTV videos their performance deteriorates. In this section, the effect of training data selection in the detector's performance is explored[7].

Building on prior deep learning work, the object detection and localization framework Faster R-CNN is employed. It combines the localization and detection tasks, while sharing convolutional layers to speed up the process. A ZF network model, pre-trained in ImageNet dataset, is selected for object detection. The model is fine-tuned to optimize the discriminative power of the features learned and therefore the detection accuracy[8]. The strategy of fine-tuning has been widely used in deep learning greatly improving the performance of a CNN. It has been shown that transfer learning, namely the use of unsupervised pre-training in a generic dataset, has significant value, offering a robust initialization of the network parameters. Two training data augmentation approaches are examined to improve fine-tuning of the examined system:

(a) the enrichment with object instances from heterogeneous sources and

(b) the addition of blurred instances of the current object collection.

In the former approach, annotated datasets featuring the examined object classes are utilized. An extended training set is created that contains samples from multiple datasets. Despite the existence of several annotated datasets, their content is produced with quality measures that are superior to the conditions that a normal CCTV system will face. Therefore, the features learned by a deep learning object detector are often plagued by many missing detections, especially in action scenes. Following the latter approach, the training set is augmented with blurred instances to enhance the robustness to motion blur.

The core of the Hungarian algorithm is to find augmented paths, it is a kind of with augmented path to calculate the maximum matching algorithm of binary chart. Recognition algorithm is divided into three steps: The first step, image preprocessing. To improve the efficiency of operation, the inspection before every image is compressed, the aspect ratio is 640 x 480, the transformation of color images to grayscale. The second step, the image input trained Faster - R - CNN network model, detect the target location. The third step, tracking moving targets. On the premise of known motion target location, the use of the Hungarian matching algorithm of moving target with labels, localization of the moving target tracking can be achieved.

CCTV cameras often have PTZ capabilities that are used by their operators to track suspicious activities in a scene. These camera operations constitute a serious challenge for object detection and tracking due to the implicit scale assumptions made. Object detection techniques have a predefined range of scales that are supported, to minimize detection errors. In this section it is proposed to dynamically adjust this scale range based on predictions of the tracked objects' size in the next frame. The first step towards this approach is to have an accurate estimation of the detected object's scale and pose. Recently, a new module was proposed that applies a spatial transformation to a feature map during a single forward pass. The spatial transformer network (SPN) [9] can be used as a new type of layer in a feed-forward convolutional network. It learns an affine transformation of the input and uses bilinear interpolation to produce its output allowing it to zoom, rotate and skew the input.

III. METHODOLOGY

The experimental setup for the validation of the above concepts is being described. Given that pedestrians are the main object class of interest, the experiments will focus on the pedestrian detection without losing its generality. For this purpose, several datasets have been selected to feature the experiments. VOC2007 is used as a generic dataset for image classification with 20 annotated classes, including the class person. The ETH dataset is also used to extend the fine-tuning dataset. It contains annotated pedestrians on a public road. Finally, a set of videos from the Metropolitan Police of London (MET) from the riots of 2011 have been also used for qualitative validation. Those videos have been offered for research purposes in the framework of the LASIE FP7 project and they are neither annotated nor publicly available.

The first set of experiments refers to the exploration of training data augmentation strategies. The Faster R-CNN object detection framework is fine-tuned with different training sets to test their effectiveness.. The training set is infused with sequences from the ETH dataset, namely "Bahnhof" sequence (~7500 object instances) labeled as [BAH] and "Sunny Day" (~1900 object instances), labeled as [SUN]. The datasets are divided in training and testing set of equal size. 50% of the training set is used for validation purposes. The evaluation of the trained models is performed on separate testing sets that include VOC testing set and an ensemble of the VOC and ETH testing sets, respectively. The results are reported in the table given below. Experiments show that the detection accuracy seems to benefit from extra training samples, even on the original VOC testing set. Subsequently, the effect of augmenting the data with motion blur is examined. Training with the VOC2007 dataset is again used as baseline.

IV. RESULT AND DISCUSSION :

In this paper our purpose is to find object detection in CCTV footage. We used RCNN algorithm to evaluate the performance of our work. So finding the object in CCTV footage is very complicated. But with use of RCNN algorithm it make simple way to find. In fig given below

Hence the Code of object detection is successfully run OpenCV and gives the result almost accurately. Below Screenshot of result is shown.

In fig 1.it detects the human object in CCTV footage.

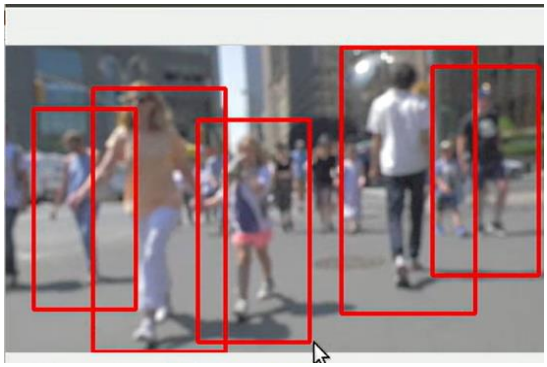

Fig. 1 Target detection from video

In Fig.2 the footage is blur but it detects the human object is very well and accurate

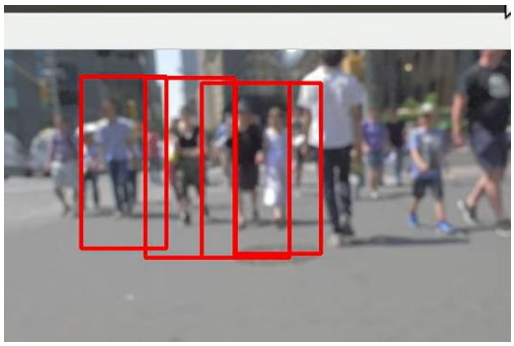

Fig. 2 Target detection from crowded scene

Fig. 3 Detection of all target with occlusion handling

In fig. 3 there the footage is very clear hence it detects the almost all human object even in case of occlusion.

V. CONCLUSION

In this paper we work on object detection in CCTV footage. Here heterogenous training data and data augmentation is used improve detection rate in CCTV footage. Methodologies to improve RCNN algorithms of object detection in CCTV. We work on challenges in footagelike blur video ,low camera qualityand other environment factors.

In our paper we can conclude that RCNN algorithm is mostly suited for detection in CCTV footage. Also we can conclude that the if there is blur or not very clear footage RCNN algorithm work very successfully and find the object in footage.

REFERENCES

[1] Anthony C Davies and Sergio A Velastin, "Progress in computational intelligence to support cctv surveillance systems," International Journal of Computing, vol. 4, no. 3, pp. 76–84, 2014.

[2] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah, "Visual tracking: An experimental survey," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 36, no. 7, pp. 1442–1468, 2014.

[3] Hanxi Li, Yi Li, and FatihPorikli, Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, November 1-5, 2014, Revised Selected Papers, Part V, chapter Robust Online Visual Tracking with a Single Convolutional Neural Network, pp. 194– 209, Springer International Publishing, Cham, 2015.

[4] Xiangzeng Zhou, Lei Xie, Peng Zhang, and Yanning Zhang, "An ensemble of deep neural networks for object tracking," in Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014, pp. 843–847

[5] J. Ding, Y. Huang, W. Liu, and K. Huang, "Severely blurred object tracking by learning deep image representations," Circuits and Systems for Video Technology, IEEE Transactions on, vol. PP, no. 99, pp. 1–1, 2015.

[6] Jialue Fan, Wei Xu, Ying Wu, and Yihong Gong, "Human tracking using convolutional neural networks," Neural Networks, IEEE Transactions on, vol. 21, no. 10, pp. 1610–1623, 2010.

[8] Li Wang, Ting Liu, Gang Wang, KapLuk Chan, and Qingxiong Yang, "Video tracking using learned hierarchical features," Image Processing, IEEE Transactions on, vol. 24, no. 4, pp. 1424–1435, 2015

[9] Yan Chen, Xiangnan Yang, Bineng Zhong, Shengnan Pan, Duansheng Chen, and Huizhen Zhang, "Cnntracker: Online discriminative object tracking via deep convolutional neural network," Applied Soft Computing, vol. 38, pp. 1088–1098, 2016.

[10] Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121