Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230335

Computer vision model's application in the current system on object detection tasks

Feilian Huang

Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

fhuang49@gatech.edu

Abstract. The implementation of object detection algorithms would be helpful to the various fields of the current time. When object detection is applied to the surveillance camera system, it will be more efficient to locate crimes or find lost kids. This paper will investigate the performance of different object detection algorithms in a real-world scenario. With experimentation, CenterNet++ outperforms YOLO and MaskRCNN, two traditional and classic object detection algorithms, on the MS COCO dataset, which concludes that CenterNet++ can ensure both accuracy and speed.

Keyword: Video Processing Systems, Computer Vision, Artificial Intelligence, Object Detection.

1. Introduction

Surveillance cameras have been widely used in the current era to record most public areas for theft identification, violence detection, and chances of explosion [1]. By looking at the recordings, people can trace back to any time given and acquire convincing proof of any malicious activities within the range of surveillance cameras. However, many obsolete systems are only able to solve problems after it has happened, and are unable to identify, for example, crime, in real-time [2]. Without video analysis algorithms like face detection or target tracking algorithms, looking at the recordings by human force can be inefficient and laborious. Imagine there is a kid lost in the amusement park, which often happens in the real world. Instead of asking staff members to check the recordings, which might be considerably slower, most parents would rather ask staff to walk around the park to look for their child. If an object detection algorithm is implemented in the modern surveillance camera system in the amusement park, staff members can find any child within one minute. An efficient video analysis algorithm will be much more powerful than ten people sitting in front of the screen and looking for suspicious activities in public areas. Therefore, the implementation of machine learning algorithms and video analysis into the camera system will become the near future with more tools implemented to make it more accurate, such as Cloud, Fog, and Edge Computing [3].

Although there are plenty of video analysis systems and models, and they are rapidly evolving in recent years, there has not been a wide usage in the real world. This paper will investigate the different usage of machine learning algorithms in various video analysis systems and identifies their advantages and disadvantages. For example, by utilizing 3D - Convolutional Neural Networks (3D-CNN), cameras can locate people and detect actions [4]; by making use of edge computing, cameras can

DOI: 10.54254/2755-2721/4/20230335

prevent violent crimes by accurately detecting actions like pushing or pulling [5]; for more complex action recognition, the usage of a Convolutional 3D (C3D) network and the Recurrent Neural Network (RNN) can play a big role in it [6]. To test the applicability of those systems in the real-world scenario, I will build a video analysis system with an appropriate object detection model and COCO dataset, a famous large-scale object detection dataset. I will choose the recently developed CenterNet++ as the model for its high speed and accuracy. It is also widely applicable to analyze most real-world photos and recorded videos.

This paper will first discuss the related work of the development of current video analysis systems and the corresponding object detection models. With a complete understanding of the most recent development, this paper will investigate further possibilities. In the next section, I will build my system with CenterNet++ with the COCO dataset, proving its applicability and comparing its speed and accuracy to the current other works. In the last part, this paper will discuss the limitations and possible future directions.

2. Literature Review

Regarding the wide use of surveillance cameras without many implementations of video analysis to the system, Kardas and Cicekli proposed a surveillance video analysis system (SVAS) in 2017. With an Interval-Based Spatial-Temporal Model (IBSTM), the system can perform automatic learning, detection, and inference of the video. Their experiment on CAVIAR Dataset and BEHAVE Dataset all show good results on object detection tasks [7]. Nevertheless, they have a comparatively simple design and have only been tested on two datasets. Later, more literature has proposed models implementing a surveillance system to fill the gap in the research on object detection or extend the usage of machine learning algorithms to other tasks. In 2018, Arunnehru, J., et al implemented the 3D-CNN model in their system to perform human action detection tasks. By accurately assigning a 3D bounding box to the person, their system shows a high accuracy towards the KTH video dataset, which covers indoor and outdoor actions such as waving, clapping the hands, running, boxing, walking, and jogging [4]. In the same year, another group of researchers used Convolutional 3D (C3D) network and the Recurrent Neural Network (RNN) to perform action recognitions on UCF101 and HMDB51 datasets [6].

Furthermore, several special techniques have been developed to improve the accuracy or speed of those models and systems. Extracting key frames from surveillance videos before video processing will significantly improve the accuracy and efficiency of video analysis [8]. Knowledge distillation can also be a powerful tool to reduce the size of large models without decreasing their accuracy [9]. Masked Feature Prediction (MaskFeat) is another powerful tool to recover the Masked region of a photo or video, which would also be helpful in a real-life scenario [10].

3. Method

With an abundance of object detection models and techniques for improvement, it is still unclear to us which model or improvement technique shows the most accurate result for real-world detection when placed on, for example, surveillance cameras. Take the road situation for example; we want the models to precisely place a bounding box on every object it has detected, such as cars, trucks, buses, and even people. Therefore, we will make use of Microsoft Common Objects in Context (MS COCO) dataset, which is taken from the real world and contains 2.5 million labeled instances in 328k images that are recognizable for even 4-year-old children. It also contains many more objects in an image (with an average of 7.7 objects per image) than ImageNet (average of 3.0) and Pascal (average of 2.3). Therefore, training the COCO dataset will benefit us by learning complex models and making precise predictions [11], and we can detect non-canonical perspectives of objects and contextual reasoning between objects and the precise 2D localization of objects [12]. This paper will train and test those computer vision models based on the same dataset, then we will partition the dataset into a training dataset, which takes 80% of the original dataset, and the other 20% will be the testing dataset. Afterward, according to the result shown in the testing dataset, this paper will investigate their advantages and disadvantages of them by assessing their accuracy, precision, robustness, etc.

DOI: 10.54254/2755-2721/4/20230335

We chose to use CenterNet++ because it was published recently and showed promising results in their paper [13]. CenterNet++ can identify objects as axis-aligned boxes in an image. By modeling every object as a single point, which will become the center point of the bounding box, CenterNet++ can make use of KeyPoint estimation to find center points of given input images and regresses to all other object properties, such as size, 3D location, orientation, and even pose [14]. Those features make CenterNet++ a robust model for object detection. In this paper, we will design an experiment testing how CenterNet++ performs in the MS COCO 2017 dataset. For comparison, we will also utilize some classic object detection models such as YOLO and Mask-RCNN. YOLO (You Only Look Once) is unified, real-time object detection, which learns the generalizable representations and reason globally in making predictions. It combines all components in the computer vision task into a single neural network, which utilizes the features of the whole image to make predictions and draw the corresponding bounding boxes [15]. This feature makes YOLO an incredibly efficient model for object detection tasks. The mask-RCNN model is also a classic object detection model that classifies object proposals using deep convolutional networks. It is an extension of the Faster-RCNN model, a simple, flexible, and general framework for object instance segmentation. Mask-RCNN extends Faster-RCNN by adding a branch, which is only a little overhead, for predicting an object mask in parallel with the existing branch for bounding box recognition. It is also a powerful and classic tool for object detection [16]. Although CenterNet++ is the most recent approach and has shown very promising results, this paper will fill the gap that they haven't been trained, tested, and compared on the MS COCO 2017 dataset, which contains data that makes the task very close to real-time object detection. In addition, this paper will make use of the techniques described in the related works to enhance the results, and we will also test how those techniques will help each model when performing their predictions.

Here, this paper will present how the pipeline of the CenterNet++ works: The hourglass-like network is its backbone, and it can fit a huge amount of backbone networks such as DarkNet-19, R-101, HG-52, etc. Then, it will apply cascade corner pooling and center pooling. Fig. 1 illustrates its basic architecture. With this method, CenterNet++ can find the corner and center of the object detected in a specific image, thus giving an accurate bounding box surrounding the object.

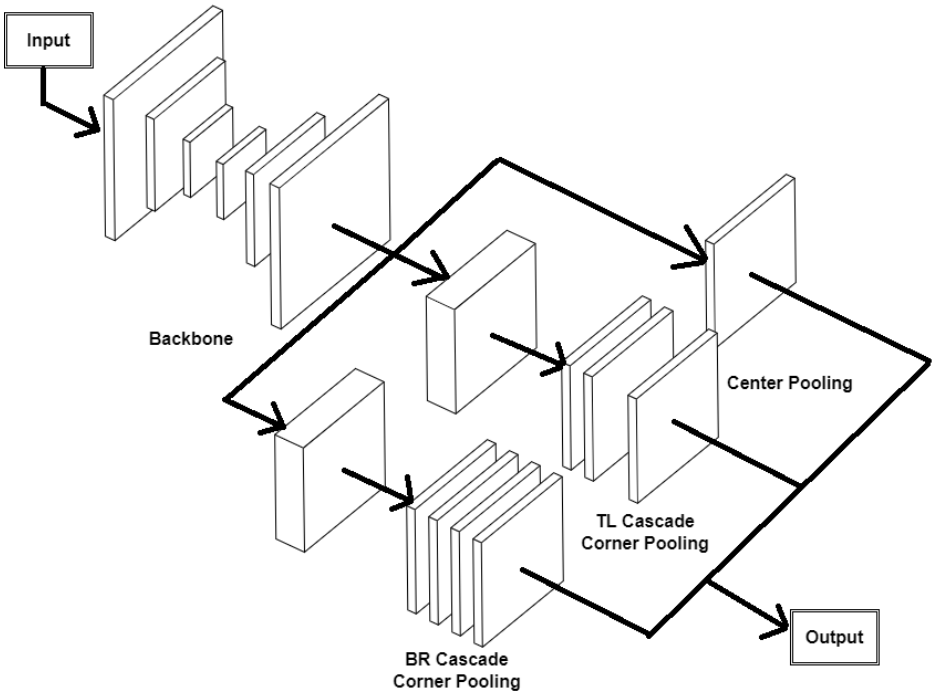

Figure 1. The architecture of CenterNet++.

DOI: 10.54254/2755-2721/4/20230335

4. Discussion

We have trained YOLO, Mask-RCNN, and CenterNet++ with the MS COCO 2017 dataset. Figure 1 shows a comparison of how different object detection algorithms are applied to the same sample image. It is an arbitrary photo taken in the street, and we visualize the result of different algorithms by drawing the bounding boxes with its label to the corresponding object in the photo. Given the information from Figure 2 and take the first row as an example, we can see that the CenterNet (on the left) performs better than YOLO (in the middle) and Mask RCNN (on the right) in ways that it accurately predicts and gives a bounding box to the major objects in the photo. While YOLO isn't powerful enough to treat situations when the objects are crowded. Also, the "traffic light" at the bottom seems to come out from nowhere. Mask RCNN seems to be easy to get confused when part of the person's body is blocked. For example, the person in the center that drives the three-wheel car was identified as two people by the Mask-RCNN model: It identifies the body as one person and gives a bounding box to his hand as another person. In addition, like YOLO, the Mask-RCNN model is easy to be confused when detecting a crowded object. The right part of the third picture is very messy are hard to visualize those bounding boxes. Therefore, with the sample image provided, it is already very clear that CenterNet++ has better performance than the other two models. In the following paragraphs, we will provide complete statistics to illustrate the reason why this paper prefers CenterNet++ over the other two classical models for object detection tasks.

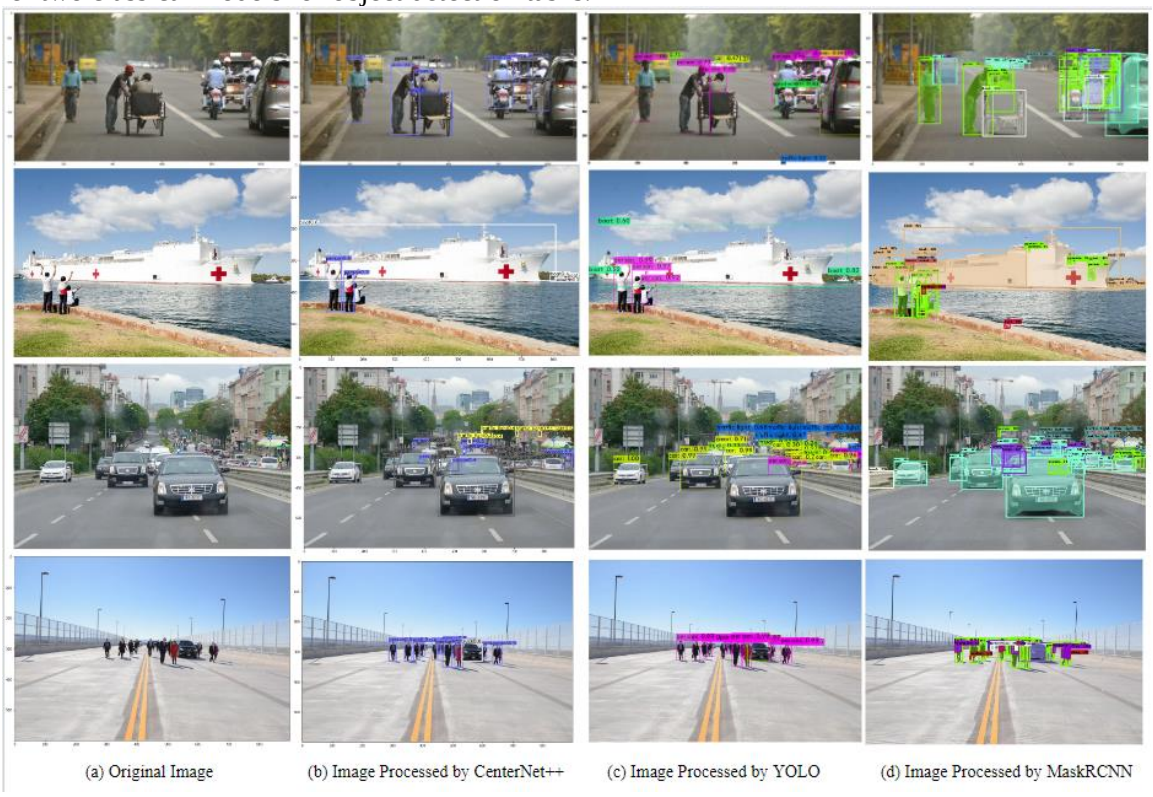

Figure 2. A comparison of the sample image and the processed image by CenterNet++, YOLO, and MaskRCNN.

CenterNet++ outperformed YOLO and MaskRCNN not just in this sample picture, but also in general. According to the experiment, we recorded the FPS and mean average precision (mAP) of their performance on the MS COCO dataset. The result is shown in Table 1. We can see that YOLO is faster than MaskRCNN, but with less accuracy; MaskRCNN is more accurate, but slower. However, CenterNet++ is way better than YOLO and MaskRCNN in both metrics, which concludes that CenterNet++ is better both in speed and accuracy [13].

DOI: 10.54254/2755-2721/4/20230335

Table 1. Performance comparison with different models on MS COCO dataset.

	FPS	mAP
CenterNet++	30.5	43.6
YOLO	20.0	33.0
MaskRCNN	11.0	39.8

In addition, CenterNet++ can be applied to a wide range of systems. Any computer system that has installed the PyTorch library can utilize it and do object detection tasks. It can be installed in the shopping mall to spot thieves at the very first moment; It can also be applied in the traffic system to automatically detect if any cars are Overspeed; Moreover, when it is used in the amusement park, no parents would need to worry about the chance of lost children anymore. In conclusion, CenterNet++ is an effective algorithm, not only guaranteeing speed and accuracy but also highly compatible with almost all systems. It will be highly beneficial to use in many real-world situations.

5. Conclusion

With all the experimental designs and testing, we found that Center++ has the best performance when testing on the COCO 2017 dataset, which highly resembles the situation in the real world. CenterNet++ is accurate and fast, making it an efficient model for object detection tasks. However, there still exist some limitations regarding the experiment procedures. First, this paper only conducts the experiment based on one dataset and three specific models. To ensure generality, it is better to do more tests on more distinct models and datasets. In addition, this paper only makes use of simple metrics to assess the power and performance of each model. Although they are most commonly and famously used in the object detection field. It is essential to include more ways to measure the performance of models. In future works, it is recommended to include more models and different datasets in the experiment. In addition, looking for more distinct ways to measure their performance will benefit us to see a bigger and much more complete picture of the tradeoffs between the popular models in the current time.

References

[1] Sreenu, G., Saleem Durai, M.A.2019 Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *J Big Data 6*, 48.

[2] Whittaker, Danielle.2021 “Why AI CCTV Is the Future of Security and Surveillance in Public Spaces.” *Security*, 14 Dec. 2021.

[3] Tsakanikas, Vassilios, and Tasos Dagiuklas.2018 “Video Surveillance Systems-Current Status and Future Trends.” *Computers & Electrical Engineering*, vol. **70**, pp. 736–753.

[4] Arunnehru, J., et al. “Human Action Recognition Using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos.” *Procedia Computer Science*, vol. 133, 2018, pp. 471–477.

[5] D., Aishwarya, and Minu R.I. 2021 “Edge Computing Based Surveillance Framework for Real Time Activity Recognition.” *ICT Express*, vol. **7**, no. **2**, 2021, pp. 182–186.

[6] Yuan, Yuan, et al. 2018“Action Recognition Using Spatial-Optical Data Organization and Sequential Learning Framework.” *Neurocomputing*, vol. **315**, 2018, pp. 221–233.

[7] Kardas, Karani, and Nihan Kesim Cicekli.2017 “SVAS: Surveillance Video Analysis System.” *Expert Systems with Applications*, vol. **89**, 2017, pp. 343–361.

[8] Wang, Dong, et al.2018 “Dairy Goat Detection Based on Faster R-CNN from Surveillance Video.” *Computers and Electronics in Agriculture*, vol. **154**, 2018, pp. 443–449.

[9] Beyer, Lucas & Zhai, Xiaohua & Royer, Amélie & Markeeva, Larisa & Anil, Rohan & Kolesnikov, Alexander. 2021. Knowledge distillation: A good teacher is patient and consistent.

DOI: 10.54254/2755-2721/4/20230335

[10] Wei, Chen & Fan, Haoqi & Xie, Saining & Wu, Chao-Yuan & Yuille, Alan & Feichtenhofer, Christoph. 2021. Masked Feature Prediction for Self-Supervised Visual Pre-Training.
[11] Lin, Tsung-Yi & Maire, Michael & Belongie, Serge & Hays, James & Perona, Pietro & Ramanan, Deva & Dollár, Piotr & Zitnick, C. 2014. Microsoft COCO: Common Objects in Context. 8693.
[12] Lin TY, Maire M, Belongie S, et al. 2014 Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
[13] Duan, Kaiwen & Bai, Song & Xie, Lingxi & Qi, Honggang & Tian, Qi. 2022 CenterNet++ for Object Detection.
[14] Xingyi Zhou, Dequan Wang, Philipp Krähenbühl. Objects as Points arXiv preprint arXiv:1904.07850
[15] Redmon, Joseph & Divvala, Santosh & Girshick, Ross & Farhadi, Ali. 2015 You Only Look Once: Unified, Real-Time Object Detection.
[16] He, Kaiming & Gkioxari, Georgia & Dollar, Piotr & Girshick, Ross. 2017 Mask R-CNN. 2980-2988.