Deep learning based target detection algorithm for motion capture applications

Haitao Wang1*, Xin Tong2 and Fengyun Lu3

1Faculty of Graduate Studies, China People's Police University, Langfang, Heibei, China
2College of Information and Cyber Security, People's Public Security University of China, Beijing, China
3College of Police Command, China People's Police University, Langfang, Heibei, China

*Corresponding Author:Haitao Wang;1270259754@qq.Com;18838210420

Email:1270259754@qq.com; tongxindotnet@outlook.com; tqbr008@163.com

Abstract: Motion capture technology is the use of external devices to perform data recording and posture reproduction of the displacement of human structures. Deep learning algorithms are playing an increasingly important role in motion capture technology as the technology involves data that can be directly understood and processed by computers in terms of dimensional measurements, positioning of objects in physical space, and orientation determination. This paper presents an application of a convolutional neural network system, YOLO-V4, in the field of motion capture. YOLO-V4 system weight files are small and do not require high hardware requirements. It can also be implemented in PyTorch so that it can be deployed on mobile devices, enabling edge devices to run these models as well, relieving the space constraint of immovable signal capture devices and providing the advantages of high accuracy and high detection rate.

1. Introduction

Motion capture technology is to set trackers in the key parts of the moving object, the motion system to capture the position of the tracker, and then through the computer algorithm processing, the three-dimensional spatial coordinates of the data. Since 1914, when the animator Max Fleischer invented the technique of Rotoscoping, the technique has been developed and widely used in animation and film. Subsequently Motion capture technology applied in gait analysis, biomechanics, ergonomics, etc. However, due to limitations in equipment, technology and algorithms, motion capture technology is often limited to fixed sites and requires complex supporting equipment, making it difficult to develop applications in areas that require a wide spatial range and strong object movement capabilities.

This thesis presents an idea for applying the current state-of-the-art YOLO-V4 in the field of target detection to motion capture. YOLO-V4, based on YOLO-V3, is modified and integrated with some new technologies to optimize the computing speed, which can achieve the accuracy of the current cutting-edge technology while maintaining a high processing frame rate. In addition, YOLO-V4 simplifies and optimizes some of the newly proposed algorithms so that they can be trained on a single
GPU and can be run on mobile devices, which, when combined with drones and other devices, can remove site constraints and increase the possibility of applying this technology in more areas.

2. Brief description of target detection algorithms and motion capture techniques

2.1. Target detection algorithms

2.1.1 Introduction to Target Detection Algorithms

Object detection, which is a machine learning algorithm designed to locate the desired object in an image or video and to detect its position and size, is one of the central problems in computer vision today. There are two main directions: two-stage algorithms such as RCNN, fast-RCNN; one-stage algorithms such as YOLO, SSD. In this paper, YOLO-V4 is one stage. The results of the test with other target price measurement systems are shown in the Figure1.

![Figure 1 Test Result](image)

2.1.2 Introduction to YOLO-V4

YOLO-V4 is a rapid target detection system designed for use in practical working environments, which has been updated with versions of YOLO-V1, YOLO-V2 (YOLO9000), and YOLO-V3. Initially, YOLO was proposed to address the problem of deep machine learning target detection speed. Compared to the traditional frameworks RCNN, fast-RCNN and faster-RCNN, YOLO solves object detection as a regression problem. The input image, after an inference, gives the position, category and corresponding confidence probability of all objects. From YOLO-V1 to YOLO-V4, issues such as detection accuracy and object localization have been continuously optimized, and small object recognition has been enhanced to reach the cutting edge of current detection systems in terms of speed and accuracy.

Following the introduction of YOLO-V4, Ultralytics has been working on "YOLO-V5" and has open-sourced the "YOLO-V5 code". As Ultralytics is not the original author of YOLO, most believe that this is just an implementation of YOLO-V4 with improved performance. While the "YOLO-V5" designation is still controversial, Ultralytics does bring enhanced real-time target detection technology to the table. Like the "YOLOv3" project, the open-source "YOLO-V5" is based entirely on PyTorch, providing the foundation for YOLO to be used in motion capture and implemented on mobile devices.

2.2. Introduction to motion capture techniques

Motion capture is a technique that uses external devices to record data and restore posture to the displacement of human structures. The more commonly used motion capture techniques are classified as acoustic, mechanical, optical, electromagnetic, and inertial navigation, and the equipment required is typically sensors, signal capture equipment, data transmission equipment, and data processing equipment. Among them, the mechanical type is low cost and high accuracy, and can do real-time measurement, but the mechanical sensor is inconvenient to wear and use, and is mostly used for static shape capture and key frame determination; Acoustic-style capture is poor in real time, with large delays and hysteresis, and average accuracy. In contrast, optical, electromagnetic and...
Principles of YOLO-V4

YOLO-V4 has five main basic components: CBM, consisting of Conv, Bn, Mish activation functions; CBL, consisting of Conv, Bn, Leaky_relu activation functions; Res unit, enabling deeper network construction; CSPX, consisting of three convolutional layers and X res unit modules; SPP, multi-scale fusion using $1 \times 1$, $5 \times 5$, $9 \times 9$, and $13 \times 13$ maximum pooling methods.

The structure of the YOLO-V4 system, like most target detection systems, is divided into four main parts, Input, BackBone, Neck and Prediction, as shown in Figure 2.

3.1. Input

The input mainly uses Mosaic data enhancement, cmBN, and SAT self-confrontation training. The Mosaic data enhancement is performed by randomly scaling, cropping and arranging the four images to enhance the robustness of the neural network, thus enriching the data set and reducing the GPU, which enables a single GPU to achieve good training results.

3.2. BackBone

3.2.1. CSPDarknet53 network

BackBone uses the CSPDarknet (Cross Stage Partial Network) 53, Mish activation function, and Dropblock. BackBone contains five CSP module, so the graphics can be reduced 32 times after BackBone to get a feature map, reducing the computational bottleneck. CSPDarknet53 enables the system to be lightweight while maintaining accuracy.

3.2.2. Mish

Compared to leaky_relu, Mish enables some improvement in network accuracy for CSPDarknet53, therefore, YOLO-V4 changed the leaky_relu in BackBone to Mish. The function is as Equation (1).

\[ y_{\text{mish}} = x \cdot \tanh(\ln(1 + e^{x})) \]  

3.2.3. Dropblock

Unlike the traditional Dropout random deletion of neurons, Dropblock censors localized areas of the image, which draws on Cutout data enhancement. While Cutout only works on the input side, Dropblock can be applied to every feature map and the probability of culling can be changed at any time.

3.3. Neck

Neck mainly uses SPP (Spatial Pyramid Pooling) modules, FPN+PAN.

3.3.1. SPP modules

Almost all convolutional neural networks require the input data to be fixed in size, but in many cases the aspect ratio of the data obtained is not fixed, and if the image is cropped, it is likely that important data will be lost. SPP divides the convolutional feature maps into several copies, pools them, and converts them into one-dimensional matrices, so that data of any size can be entered.
3. 3. 2. **FPN+PAN**

FPN+PAN learn from PANet of CVPR '18. FPN fuses top-level features by up-sampling and low-level features in a top-down direction. And YOLO-V4 adds a reverse direction feature pyramid containing two PAN structures after FPN. FPN is responsible for communicating strong semantic features, and feature pyramids are responsible for communicating strong positional features. Both perform feature aggregation on the detection layer from opposite directions, improving the feature extraction capability of YOLO-V4.

3. 4. **Prediction**

3. 4. 1. **CIOU_loss**

CIOU_loss takes the aspect ratio into account, in addition to the overlap area and the distance from the centroid, when compared to the normal objective box regression function. The function is as Equation (2).

\[
\text{CIOU_loss} = 1 - \left( \text{IOU} - \frac{\text{Distance}}{\text{Distance}_C} + \sqrt{1 - \text{IOU}^2} \right)
\]

In addition to the tricks mentioned above, YOLO-V4 optimizes the network using CmBN, WRC, DIOU_nms, etc. to get this efficient and usable model.

4. **The advantages of YOLO-V4 in motion capture applications**

4. 1. **Low hardware configuration requirements**

Deep learning generally has high requirements on hardware such as graphics cards and CPUs, the amount of data analyzed is huge, and the training process of the model is long. Using RCNN as an example, it takes several days to train 11 epochs on most models with a single GTX1080TI. If train from scratch, 350-500 epochs would take up to several months or even close to a full year. YOLO-V4 not only optimizes the computation speed, but also improves the training input to reduce the training threshold by incorporating some new technologies. A regular GTX-2080ti or Titan-XP GPU will be able to do the training. The low cost makes it easy for YOLO-V4 to be used and commercialized in the motion capture field.

4. 2. **Fast and accurate detection rate**

The YOLO-V4 was able to achieve 65 (FPS) with 43. 5% AP accuracy at launch, 10% higher accuracy and 12% faster than YOLO-V3, leading the field of computer vision. The optimized "YOLO-V5" achieves inference times as fast as 0. 007 seconds per image, 140(FPS), while maintaining detection accuracy.

4. 3. **Deployable on mobile devices**

YOLOv4 (Darknet architecture) has a weighted file size of 244MB, the optimized "YOLO-V5" is nearly 90% smaller in size and has a weight file of only 27MB, which means "YOLO-V5" can be easily deployed in embedded devices. In addition, Ultralytics optimized "YOLO-V5" is also based on PyTorch, just like the previous "YOLOv3 based on PyTorch Replication" project. As a result, with the PyTorch ecosystem built by Ultralytics, YOLO-V5 can be easily compiled to ONNX and CoreML, making it easy to deploy to mobile devices. Motion capture can be carried out in a wider area with less space limitation, and the application field will be greatly expanded.
5. Applications of YOLO-V4 in motion capture applications

5.1. Deploying to Mobile Devices with PyTorch

PyTorch, the open source deep learning framework developed by Facebook, carries The PyTorch Mobile framework in version 1.3, which supports efficient running of machine learning on edge devices, enables Python is deployed on IOS and Andriod. At the same time, Ultralytics built the PyTorch ecosystem for the YOLO-V4, which allows it to run on edge devices such as mobile phones.

5.2. Drone gimbal with signal capture device

YOLO-V4 can be built on a drone head after being deployed to a mobile device. The YOLO-V4 can be mounted on a drone head after being deployed on a mobile device. By taking advantage of the mobility and high maneuverability of drones, the space limitation of motion capture equipment can be broken. Real-time motion capture of high speed moving objects and wide range moving objects is realized. Simultaneous capture of variable speed moving objects can also be achieved using the hovering technology of multi-wing drones.

When YOLO-V4 is combined with drones, motion capture technology can be applied to many new areas such as major sports, rugby, football and more. At the same time, there will be a wide range of applications in the military, public relations and security fields: In public safety enforcement operations, traditional enforcement mostly uses surveillance video to capture signals from pedestrians and vehicles, which often leads to interruptions in surveillance dead ends. The use of drones for law enforcement can be such a problem, while enabling enforcement activities such as automated patrols and automated range searches; In military training, police training, the traditional motion capture system can only capture signals in a small area, the continuity of the movement of training personnel and the effectiveness of training combat are greatly affected. Using YOLO-V4 and drones, efficient target detection can be achieved. Trainers can ignore field constraints and the working range of signal capture equipment to conduct complete training, ensuring realistic training while collecting training data in real time with high accuracy.

Acknowledgments

The study was financially supported by Theoretical system construction and key technology research of police combat and training data processing platform of China People’s Police University (Grant No.2019zdgg009), the Key scientific research projects of colleges and universities in Henan Province (No. 20B520008) and the Scientific research project of Henan police college (HNJY-2019-33 and HNJY-2019-39).

References

[1] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: optimal speed and accuracy of object detection.
[2] Redmon, J., & Farhadi, A. (2018). Yolov3: an incremental improvement.
[3] Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., & Cho, K. (2019). Augmentation for small object detection.
[4] Misra, D. (2019). Mish: a self regularized non-monotonic neural activation function.
[5] Wang, C. Y., Liao, H. Y. M., Yeh, I. H., Wu, Y. H., Chen, P. Y., & Hsieh, J. W. (2019). Cspnet: a new backbone that can enhance learning capability of cnn.
[6] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2019). Distance-ious loss: faster and better learning for bounding box regression. arXiv.
[7] Redmon, J., & Farhadi, A. (2017). Yolo9000: better, faster, stronger. 6517-6525.
[8] Wu, Z., Nagarajan, T., Kumar, A., Rennie, S., Davis, L. S., & Grauman, K., et al. (2017). Blockdrop: dynamic inference paths in residual networks. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
[9] Huang, Z., & Wang, J. (2019). Dc-spp-yolo: dense connection and spatial pyramid pooling based yolo for object detection. arXiv.

[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You only look once: unified, real-time object detection.