Research on Pedestrian Multi Object Tracking Based on FairMot Transfer Learning

Hongjun Chen¹, Liheng Zhao¹ and Lei Ma¹*

¹ Department of Internet of things Engineering, Jincheng College of Sichuan University, Chengdu, Sichuan, 611730, China
Corresponding author’s e-mail: chj@hefuzhi.com

Abstract. Object tracking is divided into two processes: multi object detection and multi object tracking. In this paper, the multi object detection is studied firstly, and the multi object detection algorithms based on prior box anchor and anchor-free are analyzed and compared. Then in the second process of multi object tracking, multi object tracking algorithm are researched: the multi object tracking algorithm based on prior box anchor, the multi object tracking algorithm based on anchor-free.Finally, an experiment is carried out to track the pedestrians on the road by using FairMot transfer learning, and the metrics of MOTA, IDF1, IDS, MT, ML and FPS are analyzed. Experiments show that under the method of using transfer learning, MOTA can be improved by 4.2 percentage points.

1. Introduction

MOT(Multi Object Tracking) based on deep learning has become a long-term goal in the field of computer vision. The purpose is to estimate the trajectories of multiple objects of interest in the video. It has broad application prospects in the fields of action recognition, public safety, sports video analysis, elderly care and human-computer interaction[1]. At present, single target tracking (VOT / SOT), target detection, pedestrian Re-ID (Re-Identification) are very popular directions. However, the research of multiple objects in video is relatively less [2].

The common process of target tracking is as follows: first use the object detection algorithm to detect the target, and then extract the characteristics of the target through other tracking algorithms, and locate the target after the next appearance of the image. In the first step of the detection algorithm, most of the mainstream object detection algorithms can achieve a high accuracy and robustness, the main difficulty is the tracking algorithm track. On the other hand, because the object in the video is moving, it is necessary to overcome many environment changes caused by the moving of the object. Sometimes the environment and scene changes are quite complex, including the change of object speed, the noise caused by various scene mutations, the rotation of the object, and serious occlusion problems. These problems need to be overcome and solved by current object tracking algorithms. Target search is used to predict the motion state of the object in the next frame, locate the object by predicting the possible position of the target, and reduce the search range of the object [3].

2. Object detection algorithm

2.1. Detection algorithms based on prior anchor

Most of the current object detection algorithms are based on anchor box and FPN (Feature Pyramid) prior detection algorithm. When the network is training, a group of anchors with different scales and
positions can be set in advance. In order to cover all the dimensions and positions of the whole picture, each anchor box is responsible for detecting the offset of the target whose intersection and union ratio is greater than a certain threshold value. See Figure 1. Generally speaking, the function of anchor box is to scan all areas of the picture through a little rule, and to determine whether the target appears in the area, and to calculate the deviation between the anchor box and the real box of the object. The representative works of this anchor box detection algorithm are yolo-v3 and fast RCNN. This detection method is intuitive, efficient, and has good performance in speed and accuracy. But its disadvantages are also obvious. The redundancy of anchor box and a large number of parameters will increase the training complexity and slow calculation speed of the model.

Figure 1. Detection process based on prior anchor.

When the uncertainty of target motion is low, Mahalanobis distance is a good correlation measure. However, in practice, when the camera moves, Mahalanobis distance can not match a lot, which will make this measure invalid.

2.2. Detection algorithms based on anchor-free
The detection algorithm based on Anchor-free does not use a priori box, such as CenterNet. This paper focuses on the CenterNet model. CenterNet does object detection through key point estimation. CornerNet takes the two corners of box as the key points; ExtremeNet detects the top, bottom, left, right and center points of all targets; all these networks are built on the robust key point estimation network like ours. However, they all need to go through a key point grouping stage, which will reduce the overall speed of the algorithm; while CenterNet only extracts the center point of each target, and there is no need to group or post process the key points.

Make \( I \in R^{H \times W} \) as the input image, its width is \( W \) and its height is \( H \). Our goal is to generate thermal maps of key points \( \hat{Y} \in [0,1]^{R \times R \times C} \), Where \( R \) is the output stripe (i.e. the size scaling ratio), \( C \) is the number of key point types (i.e. the number of output feature map channels). The types of key points are: human joint points with \( C = 17 \), which are used for human pose estimation; \( C = 80 \) for target detection. We use \( r = 4 \) by default. \( Y_{x,y,c=1} \) represents the key points detected; \( Y_{x,y,c=0} \) represents the key points detected. For the key point \( C \) of ground truth (GT), its position is \( p/\epsilon R^2 \), We pass the GT key point through the Gaussian kernel to get the corresponding key point \( C \) of low resolution (after down sampling):

\[
Y_{x,y,c} = \exp\left(-\frac{(x-p_x)^2 + (y-p_y)^2}{2\sigma^2}\right)
\]  

(1)

Scatter it on the heat map \( Y \in [0,1]^{R \times R \times C} \), Where \( \sigma \) is the target scale adaptive standard deviation. If two Gaussian functions overlap for the same class \( c \) (the same key point or target class), we choose the one with the largest element level. Where \( p \) is the target scale adaptive standard deviation.
3. Object tracking algorithm

3.1. Object Tracking Algorithm Based on Anchor

Tracking-by-detection algorithm (See Figure 2.) is widely used in the field of MOT which is becoming more and more mainstream. Previous algorithms, such as flow network formula and probability graph model, are processing the global optimization problem of the whole process, but is not suitable for online scenes, and its target identification must be available in each time step. More traditional are hypothesis tracking (MHT) and joint probabilistic data correlation filter (JPDAF), These methods perform data association based on frame by frame. Recently, these methods have been re-recognized due to the success of the detection problem.

![Flow chart of target tracking algorithm based on anchor box.](image)

The previous SORT algorithm uses a simple Kalman filter to deal with the correlation of frame by frame data and uses the Hungarian algorithm is used to measure the correlation, and this simple algorithm achieves good performance at high frame rate, but because SORT ignores the surface features of the detected object, it is only accurate when the uncertainty of the object state estimation is low. In Deep SORT, we use more reliable measures instead of correlation measures, and use CNN network to train large pedestrian data sets to extract features, which enhances the robustness of the network to losses and obstacles.

State estimation: use an 8-dimensional space to describe the state of the trajectory at a certain time.

\[(u, v, r, h, x^*, y^*, r^*, h^*)\]  

This represents the position, aspect ratio, height of the center of the bounding box and the corresponding velocity information in the image coordinates. Then a Kalman filter is used to predict the update trajectory. The Kalman filter adopts uniform velocity model and linear observation model.

Firstly, in Fairmotor object tracker, the input image is fed to the encoder decoder network to extract the high-resolution feature image. Then, we add two simple parallel headers to predict the bounding box and Re-ID features, respectively. The feature of the center of the prediction object is extracted to link the time bounding box.

3.2. Object tracking algorithm based on anchor-free: FairMot

FairMot is object tracking algorithm based on centernet based on centerne, which is published by Huake and Microsoft Asia Research Institute on April 8, 2020. It is usually used in combination with Re-ID called CenterNet+Re-ID, the detection model first defines the boundaries of the box to locate the objects of interest. Then, the association model extracts Re-ID features for each boundary box, and
links them to one of the existing tracks according to some measures defined on the features. In recent years, significant progress has been made in object detection and Re-ID respectively, which improves tracking performance. See Figure 3.

![FairMot Sketch Map](image)

**Figure 3. FairMot Sketch Map.**

4. **Transfer learning**

Transfer learning is to apply the knowledge or pattern learned from a certain field or task to different but related fields or problems. This paper adopts transfer learning based on shared parameters, which uses common parameters or prior distribution between spatial models by the similarity between the disclosing data MOT20 and self-collected data, so as to achieve the purpose of knowledge transfer through further processing.

5. **Experiment and analysis**

5.1. **Experimental platform and data**

The hardware environment of the experiment is Intel Core i5-7300 CPU, 3.60GHz main frequency and 32GB memory, NVIDIA 1080ti 11G Graphics card. The software environment is Python, Pytorch(deep learning framework), CUDA library, CUDNN, OpenCV, Ubuntu 16.4. The dataset are the public data set MOT20 and self-collected data sets.

5.2. **Training and metrics**

The parameters of fairmot model are preloaded as the initial parameters of transfer learning. A total of 150 epochs are trained. The batch size is set to 8 and the initial lr (learning rate) is set to 0.001. When the learning rate is 0.001, the accuracy of transfer learning training method is improved rapidly. After the 50th epoch reduces the learning rate, the training accuracy of 60th epoch converges. Therefore, the use of transfer learning method is an effective way to improve the training efficiency and accuracy, especially when the amount of data is small or the performance of deep learning equipment is limited, the role of transfer learning method is more prominent, MOTA reaches 60.1%, and the accuracy rate is improved by 4.2 percentage points. The performance of FairMot model and its transfer learning model were analyzed from Mota, IDF1, IDS, MT, ML, FPS.

![Use FairMot transfer learning.](image)

**Figure 4. Use FairMot transfer learning.**

|                  | MOTA  | IDF1  | IDS   | MT     | ML    | FPS  |
|------------------|-------|-------|-------|--------|-------|------|
| MOT20            | 58.7  | 63.7  | 6013  | 66.3%  | 8.5%  | 13.2 |
| OurData          | 61.17 | 65.9  | 7080  | 69.2%  | 8.1%  | 30   |

MOTA (Multiple Object Tracking Accuracy). MOTA considers the false detection, missed detection and ID exchange in the trajectory.
\[ MOTA = 1 - \frac{\sum (FN_i + FP_i + IDSW_i)}{\sum GT_i} \quad (3) \]

Among them, FN is False Negative, FP is False Positive, IDSW is ID Switch, GT is the number of ground truth objects. MOTA considers the matching errors in all frames in tracking by FN, FP and ID switch. MOTA gives a very intuitive measure of the performance of the tracker in detecting objects and keeping tracks, which is independent of the accuracy of object position estimation. The value of MOTA should be less than 100. When the error generated by the tracker exceeds the object in the scene, the MOTA will be negative[4].

IDF1 is a comprehensive consideration of ID accuracy and ID recall.

\[ IDF1 = \frac{2 \text{IDTP}}{2 \text{IDTP} + \text{IDFP} + \text{IDFN}} \quad (4) \]

It is worth noting that, different from Mota, the indicators of TP, FP and FN in IDF1 take ID information into account, while only IDSW item in MOTA takes ID information into account. From this point of view, IDF1 is more sensitive to the accuracy of ID information in trajectory.

MT (Mostly Tracked). Meet the requirement that the ground truth matches the successful track at least 80% of the time, which accounts for the proportion of all tracking targets. Note that MT and ml here have nothing to do with whether the ID of the current track changes, as long as the ground truth matches the target.

ML (Mostly Lost). This means the percentage of all tracked targets that meet Ground Truth and matches successfully in less than 20% of the time.

ID Switch. The number of times that Ground Truth assigned ID has changed.

The result to use FairMot transfer learning model, seeing Figure 4.

6. Summary
In this paper, FairMot algorithm is used for transfer learning, and the performance of FairMot model and FairMot transfer learning model is compared. The MOTA performance of transfer learning is 4.2% higher than that of FairMot model. However, due to the deformation and position change caused by the change of the target's attitude, camera occlusion, large background difference and the influence of external light weather[5], there are still many difficulties to be overcome.

Acknowledgments
Our thanks to: Fund of Project of Science and Technology Department of Sichuan Province, No.19YYJC2411 and Fund of Project of Education Department of Sichuan Province, No.18ZB0502

References
[1] Liu X.M., Chen X.N.(2020)Multi target tracking algorithm based on Yolo detection. Computer engineering and Science, 4:1007-1013
[2] Bewley A., Ge Z.Y., Ott L., Ramos F., Upcroft B. (2016) Simple online and realtime tracking. In 2016 IEEE International Conference on Image Processing (ICIP), IEEE: 3464–3468
[3] Guan H., Xue X.Y., An Z.Y. (2017) Survey of online single target video tracking algorithms, Minicomputer system. 38(1):147-153
[4] Zhang Y.F., Wang C.Y., Wang X.G., Zeng W.J., Liu W.Y. (2004) A Simple Baseline for Multi-Object Tracking. arXiv technical report ,arXiv 2004.01888
[5] Yang H., Shao L., Zheng F. (2011) Recent advances and trends in visual tracking. A review neuro computing, 74(18):3823-3831