一种多模融合的实时多目标跟踪算法

徐煜秾，叶平，张治广

（北京邮电大学人工智能学院，北京 100876）

摘要：近年来，多目标跟踪已经成为计算机视觉领域的一个重要研究分支。当前，主流的多目标跟踪算法都采用模块分离的方式，然而这种模块之间相互独立的方式并不适用于现实中的实时场景。最近，一种联合检测与跟踪模块的多目标跟踪方案逐渐引起研究人员的关注。但是，简单的融合两个模块并不能取得显著的效果。针对当前对目标跟踪算法模型繁杂，性能较差的情况，本文设计改进了一种融合多个模块的多目标跟踪方法。该方法通过融合目标检测、运动模型和重识别特征提取等模块，实现多目标跟踪系统可以实时运行并且取得良好性能。该方法在 2DMOT15 数据集中取得 58.8%的跟踪准确度和 28 的运行帧率，在 MOT16 数据集上取得 70.2%的跟踪准确度以及 24 的运行帧率。

关键词：多目标跟踪；目标检测；行人重识别

中图分类号：TP391.41

An Online Multi-Object Tracking Benchmark for Real Time Tracking

XU Yunong, YE Ping, ZHANG Zhiguang

(Artificial Intelligence School, Beijing University of Posts and Telecommunications,Beijing 100876)

Abstract: In recent years, the Multi-Object Tracking (MOT) problem has become an important research branch in computer vision. Nowadays, most current MOT approaches follow the tracking-by-detection paradigm to separately conduct object detection, feature extraction, and data association. However, this separated way is not suitable for actual industrial applications in terms of operational efficiency. Recently, it has gradually become a trend for joint detection and tracking into one framework. However, merely integrating two models does not significantly improve the tracking performance. This paper designs an improved end-to-end tracking architecture that combines multiple sub-modules and can run in real-time. Our method achieves 58.8% MOTA on 2DMOT15 at 28FPS and 70.2% MOTA on MOT16 at 24FPS.

Key words: Multi-Object Tracking; Object Detection; Re-Identification

0 引言

多目标跟踪[1]-[2]（简称：MOT）已经成为计算机视觉领域的一个重要研究方向。多目标跟踪旨在一个视频或一段连续图像序列中找到运动的物体，然后将这些运动的物体在不同帧中一一对应，最后给出不同物体的运动轨迹。多目标跟踪技术在视频分析、智能安全等诸多领域具有重要的研究价值。

当前，主流的多目标跟踪算法大都采用基于检测的方式[3]（Tracking-by-Detection，TBD），即在每一帧先进行目标检测，然后再利用目标检测的结果来进行目标跟踪——这一步也叫作数据关联。基于检测的多目标跟踪算法都把目标检测和数据关联作为独立的模

作者简介：徐煜秾(1993-)，男，硕士研究生，主要从事深度学习和目标跟踪方向研究

通信联系人：叶平(1979-)，男，副教授、硕导，主要从事移动机器人视觉 SLAM 和深度学习方向研究. E-mail: yeping@bupt.edu.cn

块分开进行，这样做虽然可以使系统结构变得简单，但是每一帧都进行数据关联计算会带来计算冗余，影响实时性。最近，有部分研究者提出联合检测跟踪框架，利用目标在相邻帧之间的上下文信息，在检测的同时完成对目标的跟踪。

基于联合检测跟踪框架，本文提出一种多模融合的多目标跟踪算法，不仅融合了检测与跟踪模块减少计算冗余，并且还集成了运动模型和行人重识别特征来优化数据关联过程，使得多目标跟踪系统既能满足实时性需求，同时还取得良好跟踪精度，本文所提模型结构如图 1 所示。本文的主要贡献总结如下：

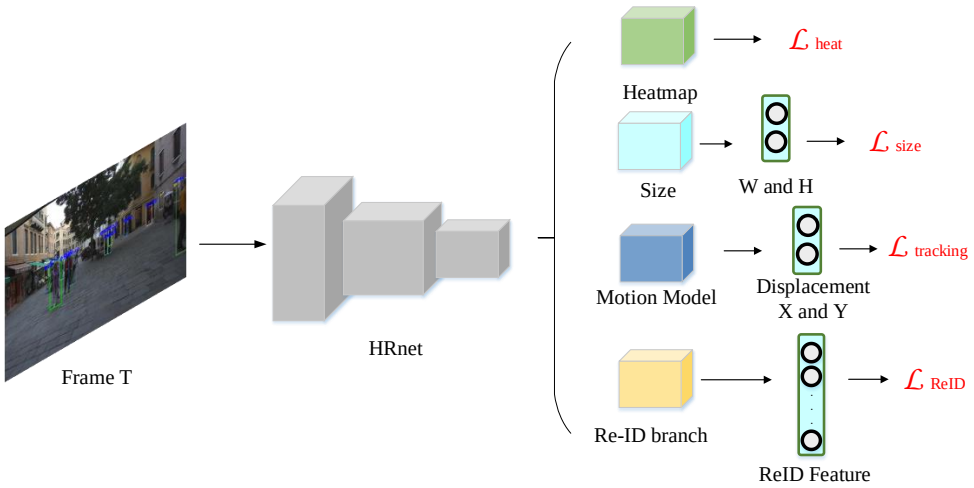

图 1　多模融合网络结构示意图

Fig. 1　Multi-Model Fusion MOT Network Structure

— 设计了一种多模融合的高效多目标跟踪模型，把目标检测、重识别特征提取和运动模型融合到一个网络。

— 使用神经网络预测模型代替传统卡尔曼滤波模型，实现目标的运动预测。

— 基于高效多模融合模型，本方法在 2DMOT15、MOT16 和 MOT17 数据集上均取得优异的实验结果。

1　相关工作

1.1　基于检测的多目标跟踪

当前流行的多目标跟踪方法大多遵循检测跟踪框架，这种方式除了需要对每帧图像进行目标检测之外，还需要执行一次数据关联过程来更新跟踪器。SORT[4]使用卡尔曼滤波器预测目标框位置，并使用二分匹配将检测框与预测框做关联。仅仅依靠预测位置在复杂场景下并不能取得良好效果，Wojke 等[5]在次基础上利用卷积神经网络提取目标特征来改善数据关联过程得到 DeepSORT。在过去，许多科研人员都专注于提高数据关联部分的鲁棒性。

这种模块间相互独立的多目标跟踪方法[6]的优点是结构简单，并且由于各个模块都相对独立，这种方法具有良好的可扩展性。此外，还可以针对每个任务分别使用最合适的模型。但是，这种方法通常运行效率较低，无法做到实时，因为每个检测框都加入到相邻帧之间数据关联[7]的计算过程中，造成计算冗余。

1.2 联合检测跟踪算法框架

最近，有部分研究者提高一种联合检测跟踪算法框架，利用目标在连续帧中上下文关系，在完成目标检测的同时完成跟踪。Feichtenhofer 等[8]使用当前帧和上一帧作为 Siamese 网络的输入，并预测目标框之间的帧间位移来完成目标跟踪。Long 等[9]仅使用 R-FCN 进一步对观测框的前景和背景进行分类，用于分类和过滤目标框。Wang 等[10]通过把重识别网络嵌入到目标检测网络中实现集成学习的方法来优化数据关联阶段。Bergmann 等[11]提出 Tracktor 方法，利用 Faster-RCNN 这类二阶段目标检测模型的特点，用上一帧的跟踪结果作为 RPN 子网络的输入，回归计算得出下一帧的跟踪结果。

尽管这些以 Tracktor 为代表的联合检测跟踪框架能够简化计算，实现多目标跟踪的实时运行，但是仅通过上一帧的回归结果并不能准确跟踪目标，尤其是在行人重叠的稠密场景。在复杂环境下，单一依靠运动模型或者重识别特征来完成数据关联并不能取得良好跟踪效果，会出现如图 2 所示的问题。但是采用常规的卡尔曼滤波做运动预测或者添加独立的重识别特征提取网络又会带来额外的计算开销，影响系统性能。针对以上问题，本文提出一种融合了目标检测、行人重识别和运动预测的多模融合网络模型，通过共享权重的方法减少计算冗余，提高运行效率，使得多目标跟踪网络既有良好跟踪精度，又能实时运行。

图 2 现有跟踪算法在复杂场景下发生误匹配的情况。在图中，蓝色目标框原本属于一位男士，使用常规方法进行跟踪之后，蓝色目标框传递到了一位女士身上，发生了 ID 跳变的情况。

Fig. 2 The existing tracking algorithms have mismatches in complex scenarios. In the picture, the blue target box originally belonged to a man. After the conventional method was used for tracking, the blue target box was passed to a woman, and the ID jump occurred.

2 多模融合多目标跟踪算法

2.1 骨干网络选取

在目标检测任务当中，高分辨率特征图因其拥有更丰富的语义信息，从而十分重要。现有的网络都是首先将输入图像编码为低分辨率表示，然后从编码的低分辨率表示中恢复高分辨率表示。然而，这种非并行的会降低计算性能，影响网络效率。为了解决计算性能低的问题，我们采用一种深度高分辨率网络[12]——HRNet。

HRNet 网络结构采用平行结构融合了多层高低分辨特征图，其网络结构如图 3 所示。通过高分辨率特征图和低分辨率特征图的融合，HRNet 拥有更好的特征表达能力。假设输入图像的大小为$W \times H$，输出特征图的大小为$C \times W_i \times H_i$，其中$H_i = H/4$和$W_i = W/4$，C为输出通道数。同样，其他高低特征融合网络也可作为骨干网络，比如 FPN 结构，BiFPN[13]结

构等。

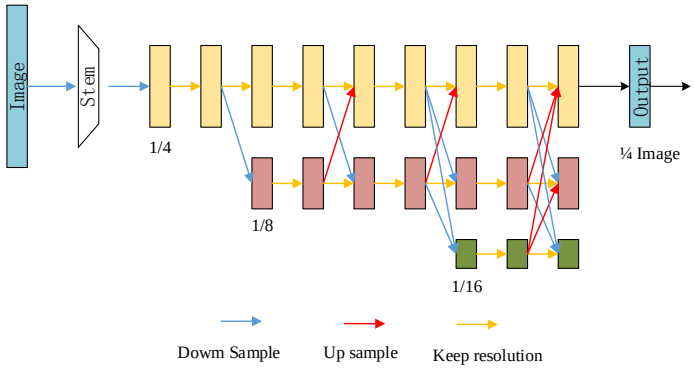

图 3 HRNet 骨干网络结构图

Fig. 3 HRNet Backbone Network Structure Diagram

2.2 目标检测分支

我们使用 CenterNet[14]网络作为我们的基础检测器。CenterNet 网络的输入为一张单独图片 $I \in \mathbb{R}^{W\times H\times 3}$，对于每个检测类别 $c \in \{0,\dots,C-1\}$，都会输出与之相对应的检测结果 $\{(\mathrm{p}_i,\mathrm{s}_i)\}_{i=0}^{N-1}$。CenterNet 识别每个目标都是通过检测它们的中心点 $\mathrm{p} \in \mathbb{R}^2$，至于检测框大小则是通过额外的两个通道 $s \in \mathbb{R}^2$ 回归得出。总的来说，我们的检测器可分为两个分支，一个用来预测目标中心点，另一个用来回归目标框大小。

2.2.1 中心点预测

给定一张图片 $I \in \mathbb{R}^{W\times H\times 3}$ 作为输入，CenterNet 输出一张用于预测目标关键点的热力图 $Y \in [0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$，称作 HeatMap。其中$R$是缩放尺度，$C$是检测类别数。在 MOT 任务中，由于只对行人做跟踪，所以类别数$C=1$。热力图上值的范围为 0-1，热力图的峰值点就是我们检测目标的中心点。给定一张图片并附带一系列标签$\{(\mathrm{p}_0,\mathrm{p}_1,\dots)\}$，使用如下 FocalLoss 损失函数[15]训练网络：

$$L_{heat}=\frac{-1}{N}\sum_{xyc}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}\log(\hat{Y}_{xyc}) & \text{if } Y_{xyc}=1\\ (1-\hat{Y}_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha}\log(1-\hat{Y}_{xyc}) & \text{other}\end{cases} \tag{1}$$

其中，$Y \in [0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$是对应带标签真值的热力图，$N$表示目标数量，$\alpha=2,\beta=4$是损失函数的超参数。

2.2.2 检测框大小回归

假设 $(x_1^{(k)},y_1^{(k)},x_2^{(k)},y_2^{(k)})$ 用来表示一个目标框的真值，那么与之相对应的中心点位置真值为 $(\frac{x_1^{(k)}+x_2^{(k)}}{2},\frac{y_1^{(k)}+y_2^{(k)}}{2})$，我们通过这两个标签值计算得到目标框大小$(x_2^{(k)}-x_1^{(k)},y_2^{(k)}-y_1^{(k)})$。假设目标框的预测大小为$\hat{S} \in \mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$，我们使用$l_1$损失函数来训练模型。

$$L_{\text{size}}=\frac{1}{N}\sum_{k=1}^{N}\left\|\hat{S}_k-s_k\right\| \tag{2}$$

2.3 模型分支

在人多的复杂场景或相机低帧率情况下，我们需要额外的运动模型来获取更准确的位置

信息。最常见的运动模型是卡尔曼滤波预测，但是卡尔曼滤波[16]需要计算额外的均值和协方差矩阵，随着跟踪数量的增加会带来巨大的时间开销。因此，我们设计了一种端到端的运动预测模型来预估目标预测位置。

我们把运动预测模型嵌入到检测网络中去，只需增加额外的两个输出通道用来获取目标在 X 方向与 Y 方向的位移。这种采用共享权重的方式将会极大得提高模型计算效率。

运动模型分支的输出是一个二维向量，分别代表目标中心点在 X 方向和 Y 方向的位移。对于每个检测目标，假定其在当前帧中的位置为$\hat{p}^{(t)}$，在上一帧中的位置为$\hat{p}^{(t-1)}$，则其位移可以定义成 $\hat{d}^{(t)} = \hat{p}^{(t)} - \hat{p}^{(t-1)}$。并把该目标的预测位移定义为$\widehat{D}_{\hat{p}(t)}^{(t)}$，然后使用$l_1$损失函数计算更新网络：

$$L_{\text{tracking}} = \frac{1}{M}\sum_{i=1}^{M}\left|\widehat{D}p_i^{(t)} - \left(p_i^{(t-1)} - p_i^{(t)}\right)\right| \tag{3}$$

其中$p_i^{(t-1)}$ 和 $p_i^{(t)}$代表目标位置的实际值。在实际训练中，我们随机选取相邻 1-3 帧作为训练数据。ReID 分支如下图 4 所示。

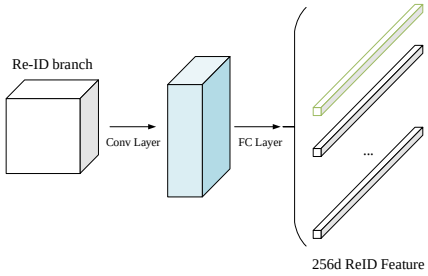

图 4　Re-ID 分支结构图

Fig. 4　Re-ID Branch Structure

2.4　重识别特征提取分支

在行人数目较多的复杂场景，仅依靠位置预测无法达到很好的跟踪效果，因此还需要结合行人重识别技术来提高数据关联成功率。行人重识别（Re-ID）技术旨在判断两个不同视角的行人是否属于同一个人，具体来讲就是提取每个目标的特征向量，并根据特征向量的相似度判断他们是否是同一个目标。

在 MOT 任务中，我们要求 Re-ID 分支所提取的特征向量具有足够的区分度。一般而言，高维度的特征向量总是有很强的判别能力，但这通常也会带来计算开销的倍增以及过度适合训练的风险。为了在性能和效率之间取得平衡，我们需要选取合适的 Re-ID 特征向量维度。通过实验对比，我们发现 256 维的 Re-ID 功能可以在精度与速度间获得更好的平衡。Re-ID 特征提取分支由几个卷积层组成，并与骨干网络相连，网络结构如图 4 所示。

在网络训练部分，我们把 Re-ID 特征提取部分当作是一个分类任务。在训练过程中，我们把训练集中具有相同身份标签的对象视为一个类。通过一个全连接层，所提取的特征向量定义为V_i^t，并学习将其映射到一个类别分布向量 $p(k)$。假设目标身份的标签值为$L^i(k)$，我们使用如下损失函数训练 Re-ID 分支：

$$L_{\text{Re}-ID} = -\sum_{i=1}^{M}\sum_{k=1}^{K} L^i(k)\log\left(p(k)\right) \tag{4}$$

其中，K是目标类别数量。

2.5 数据关联

在多目标跟踪任务中，数据关联是一个至关重要的环节，直接影响跟踪准确度与性能。数据关联问题可以看作是目标的身份 ID 在相邻帧间的传递问题。在简单情况下，通过前后帧位置可以直接完成对目标的跟踪。但是在复杂情况下，既需要 Re-ID 特征，还需要运动模型这些额外信息来获取准确的数据关联。

数据关联是一个计算繁杂且耗时步骤，具体做法就是先计算当前所有跟踪目标和下一帧的所有检测结果之间的代价矩阵，然后通过匈牙利算法[17]得出最佳匹配结果。数据关联过程会随着跟踪目标数量的增加而呈现指数级的增长。为了加速数据关联过程，我们会先把相邻帧中无重叠的跟踪目标直接赋予相应的 ID 号，以此减少后续关联计算的数量。数据关联流程如图 5 所示。

为了提高数据关联的准确率，我们使用了 Re-ID 特征代价与距离代价的加权代价矩阵作为最后的计算代价。通过实验结果，我们使用这种加权的代价矩阵在 MOT 数据集上取得优秀的结果。

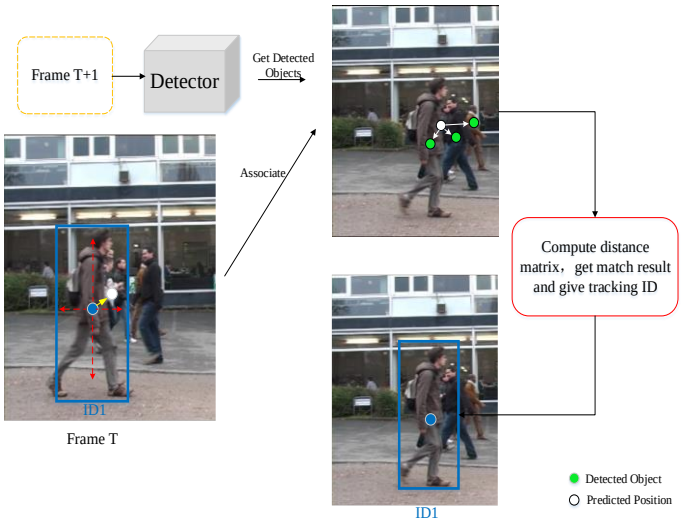

图 5 数据关联流程图

Fig. 5 Data Association Flowchart

3 实验结果与分析

我们的 MOT 模型使用了基于 Python 语言编写的 Pytorch 框架，实验结果均在 MOT 竞赛官方提供的数据集下训练和测试，包括三个主要的数据集：2DMOT15[18]、MOT16[19]和 MOT17[19]。除此之外，我们还和 MOT 竞赛榜单上的靠前方法做了全面的对比。

3.1 实验细节

3.1.1 训练部分

在实验过程中，我们的多目标跟踪器使用 CenterNet 检测器为基础，同时前端骨干网络使用 HRNet 网络。骨干网络的预训练模型是官方提供的 Image-Net 版本，并且所有实验使用的数据集均来自于 MOT challenge 竞赛官方提供的数据集。训练过程中，使用 Adam 优化器，并把训练学习率设置为 1e-4，Batchsize 设为 8。训练策略使用学习率逐渐衰减的方式，每隔 10 个周期减小十倍，总训练周期设为 30。实验所使用的 CPU 型号为 Core i9，GPU 型

号为 RTX 2080Ti。

3.1.2 跟踪部分

正如上一节所描述的那样，我们的 MOT 系统在运行过程中还有一些超参数需要设定。首先是目标检测的阈值，在 MOT15 数据集上设为 0.35，而在 MOT16 和 MOT17 数据集上设为 0.45。然后是数据关联代价矩阵的权重设置，在实验中 Re-ID 特征代价的权重设为 0.9，运动模型代价的权重设为 0.1。最后就是 IoU 阈值设为 0.5，只有满足 IoU 的目标框，才能作为数据关联对象。

3.2 评价指标与实验对比

3.2.1 多目标跟踪的评价指标

我们使用 MOT 官方给出评价指标作为实验分析参照。多目标跟踪最主要的一个评价指标就是跟踪准确度——MOTA[20]（Multi Object Tracking Accuracy），计算公式如下：

$$MOTA = 1 - \frac{\sum_t (FP_t + FN_t + IDSW_t)}{\sum_t GT_t} \tag{5}$$

其中FP_t, FN_t, GT_t和$IDSW_t$分别表示在第t帧时的误分正样本数，误分负样本数，真值框和 ID 跳变中数。

3.2.2 实验结果对比

我们的实验结果均在 2DMOT15、MOT16 和 MOT17 三个公开数据集上训练，测试结果都来自于 MOT 官方提交结果。

3.3 结果分析

在本节中，我们通过精心设计几种基准方法，对我们方法中的两个关键因素进行了严格的研究，包括骨干网络选取和 Re-ID 特征维度。

3.3.1 骨干网络对比

作为深度卷积网络中最重要的部分，直接影响了网络模型的性能。为了获取最佳跟踪效果，实验测试了几组当前主流的骨干网络和优化方法。比如 ResNet34[21]、ResNet50 和 FPN[22]。多有实验结果在表一种给出。从实验数据可以看出，ResNet50 的效果要好于 ResNet34，也符合了网路深度越深，特征提取能力越强的常规认知。当采用了 FPN 特征金字塔之后，也可以发现 MOTA 效果得到了不小的提升，这也表明了特征融合的重要性。最后采用的 HRNet 网络因为使用平行交叉特征融合结构，其效果也按预期一样优于 FPN。常见骨干网络效果对比结果在表一中给出。

3.3.2 Re-ID 特征维度实验

重识别特征提取分支因为由几个全连接层组成，最后会输出一个特征向量。高维度的 Re-ID 特征会拥有更丰富的语义信息，但是采用高维度的特征向量会增加特征代价矩阵时间开销。因此，需要选取合适的维度以达到在精度和速度间获取平衡。

常见的 Re-ID 特征大都采用 128 和 256 两种特征向量维度，在保持骨干网络不变的情况下，实验测试了 128 和 256 两种维度的特征向量对跟踪结果的影响，实验结果在表二中给

表 1　不同骨干网络对比实验结果

Tab. 1 Comparison of Experimental Results of Different Backbone

Backbone	*MOTA*	*IDF1*	*MT*	*ML*	*IDSW*
		2DMOT15			
ResNet34	52.8	52.3	29.1%	19.5%	1089
ResNet50	54.6	55.6	37.8%	17.4%	1025
ResNet34+FPN	57.6	58.2	44.5%	15.6%	986
HRNet	58.8	63.4	44.2%	12.1%	939
		MOT17			
ResNet34	61.6	65.1	37.5%	25.4%	3684
ResNet50	62.3	68.2	38.9%	24.2%	3601
ResNet34+FPN	65.8	71.4	41.2%	21.4%	3464
HRNet	72.2	72.3	38.5%	21.3%	2199

出。从表格数据也可以看到采用了 256 维的 Re-ID 特征向量，其 MOTA 指标达到最高，但是运行速度仅损失一点。

表 2 不同 Re-ID 特征维度实验结果

Tab. 2 Compare The RE-ID Feature with Different Dimensions

Re-ID 维度	*MOTA*	*IDF1*	*IDWS*	*FPS*
		2DMOT15		
128-d	56.7	59.5	996	**29.3**
256-d	**58.8**	63.4	939	28.4
		MOT17		
128-d	71.1	70.2	2287	**24.6**
256-d	**72.2**	72.3	2199	23.8

3.3.3 综合实验对比

我们最终采取的网络结构是 HRNet 作为骨干网络，Re-ID 特征维度设置为 256，并且附加运动模型分支。在 MOT 竞赛榜单上与主流的方法做了全面的对比，实验性能对比结构在表三中给出，具体跟踪效果如图 6 所示。

图 6 跟踪效果展示

Fig. 6 Show Tracking Result

表 3 与当前主流跟踪算法对比

Tab. 3 Comparisons of Tracking Results with Other Methods

Backbone	*MOTA*	*IDF1*	*MT*	*ML*	*IDSW*
2DMOT15					
AMIR[23]	37.6	46.0	15.8%	26.8%	1026
MPNTrack[24]	51.5	58.6	31.2%	25.9%	375
Tracktor[10]	46.6	47.6	18.2%	27.9%	1290
TubeTK[25]	58.4	53.1	39.3%	18.0%	854
Ours	58.8	63.4	44.2%	12.1%	939
MOT16					
MPNTrack[24]	55.9	59.9	26.0%	35.6%	431
HCC[26]	49.3	50.7	17.8%	39.9%	391
Tracktor[10]	54.4	52.5	19.0%	36.9%	682
JDE[9]	64.4	55.8	35.4%	20.0%	1544
CenterTrack[27]	69.6	60.7	-	-	2124
Ours	70.2	69.5	39.2%	17.2%	534
MOT17					
LSST[28]	54.7	62.3	20.4%	40.1%	1243
FAMNet[29]	52.0	487.7	19.1%	33.4%	3072
Tracktor[10]	53.5	52.3	19.5%	36.6%	2072
CTTrackPub	61.5	59.6	26.4%	31.9%	2583
CenterTrack[27]	67.8	64.7	34.6%	24.6%	3039
Ours	72.2	72.3	39.2%	17.2%	2560

4 结论

通过对当前主流多目标跟踪系统的学习与分析，我们发现庞大的网络模型和复杂的跟踪算法是影响多目标跟踪系统实时性的主要原因。在本论文中，我们提出一种新的多模融合的多目标跟踪模型，该模型摒弃了冗余的计算分支，利用集成学习的方法提高网络效率。具体来说，我们使用高效的骨干网络，并使用网络预测的方法代替了传统的卡尔曼滤波方法。通过一系列的改进和优化，我们的跟踪方法获取优异的结果并且可达到实时运行的速度。同时，与当前主流的方法进行比较，在 MOT 公开数据集上进行的大量实验证明了我们方法的优越性。从表三的实验对比结果中可以看出，我们的方法在性能和速度上都取得了优异的效果，并且在某些指标上取得了最佳的效果。

[参考文献] (References)

[1] S. Tang, M. Andriluka, B. Andres, and B. Schiele,"Multiple people tracking by lifted multicut and person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3701-3710.
[2] 徐涛，马克，刘才华. 基于深度学习的行人多目标跟踪方法综述[J]. 吉林大学学报(工学版)(). doi:10.13229.
[3] 尉晨阳，杨大为，张宇堃. 一种基于检测的实时在线多目标跟踪方法. 微处理机 41.06(2020):53-57. doi:.
[4] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 3645-3649, doi: 10.1109/ICIP.2017.8296962.
[5] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft. Simple online and realtime tracking. In: ICIP (2016)
[6] 茅正冲，沈雪松. 基于多特征融合的相关滤波跟踪算法[J]. 计算机与数字工程(11),2020,2645-2648+2782.

doi:.
[7] 龚轩，乐孜纯，王慧，武玉坤. 多目标跟踪中的数据关联技术综述[J]. 计算机科学,2020,47(10):136-144.
[8] C. Feichtenhofer, A. Pinz, A. Zisserman, Detect to track and track to detect. In: ICCV(2017)
[9] L. Chen, H. Ai, Z. Zhuang and C. Shang, "Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification," 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, 2018, pp. 1-6, doi: 10.1109/ICME.2018.8486597.
[10] Z. Wang, L. Zheng, Y. Liu,S. Wang, "Towards real-time multi-object tracking,"2019, [Online]. Available: https://arxiv.org/abs/1909.12605
[11] P. Bergmann, T. Meinhardt and L. Leal-Taixé, "Tracking Without Bells and Whistles," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 941-951, doi: 10.1109/ICCV.2019.00103.
[12] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang and L. Zhang, "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 5385-5394, doi: 10.1109/CVPR42600.2020.00543.
[13] M. Tan, R. Pang, EfficientDet: Scalable and Efficient Object Detection, 2019, [Online]. Available: https://arxiv.org/abs/1911.09070
[14] X. Zhou, D. Wang, and P. Krahenbuhl, "Objects as points," 2019, [Online]. Available: https://arxiv.org/pdf/1904.07850
[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, and Piotr. Dollar "Focal Loss for Dense Object Detection." 2017 IEEE International Conference on Computer Vision (ICCV), vol. PP, pp. 2999-3007. 2017.
[16] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Journal of Basic Engineering, vol. 82, no. Series D, pp. 35-45, 1960.
[17] Harold William Kuhn and Bryn Yaw. "The hungarian method for the assignment problem," Naval research logistics quarterly, pages 83-97, 1955.
[18] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. (2015). "Motchallenge 2015: Towards a benchmark for multi-target tracking,"[Online]. Available:https://arxiv.org/abs/1504.01942
[19] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. (2016). "Mot16: A benchmark for multi-object tracking," [Online]. Available: https://arxiv.org/abs/1603.00831
[20] L. Leal-Taixé, A. Milan, K. Schindler, and D. Cremers, "Tracking the Trackers: An Analysis of the State of the Art in Multiple Object Tracking," [Online]. Available: https://arxiv.org/abs/1704.02781
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770-778
[22] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117-2125.
[23] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 300-311.
[24] X. Jiang, P. Li, Y. Li, and X. Zhen, "Graph Neural Based End-to-end Data Association Framework for Online Multiple-Object Tracking," 2019, [Online]. Available: https://arxiv.org/abs/1907.05315
[25] B. Pang, Y. Li, Y Zhang, M. Li and C. Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6308-6318
[26] L. Ma, S. Tang, M. J. Black, and L. V. Gool, "Customized Multi-Person Tracker," 2019, [Online]. Available: https://is.mpg.de/uploadsfile/attachment/attachment/469/0509.pdf
[27] X. Zhou, V. Koltun, and P. Krahenb, "Tracking objects as points," 2020, [Online]. Available: https://arxiv.org/abs/2004.01177
[28] W. Feng, Z. Hu, W. Wu, J. Yan, and W. Ouyang," Multi-Object Tracking with Multiple Cues and Switcher-Aware Classification," 2019, [Online]. Available: https://arxiv.org/abs/1901.06129
[29] P. Chu and H. Ling, "FAMNet: Joint Learning of Feature, Affinity and Multi-Dimensional Assignment for Online Multiple Object Tracking," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 6171-6180, doi: 10.1109/ICCV.2019.00627.