

基于生成对抗网络的多目标行人跟踪算法

魏颖¹, 徐楚翘¹, 刁兆富¹, 李伯群²

(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819; 2. 辽宁科技大学 电子与信息工程学院, 辽宁 鞍山 114051)

摘 要: 多目标跟踪领域中,在背景复杂、目标遮挡、目标尺度和姿态变换等情况下,容易出现目标丢失、身份交换和跳变等问题。针对这些问题,提出了一种基于检测的多目标跟踪算法,使用改进的YOLO人体人脸关联算法,对当前帧待检目标进行分类和位置检测,使用生成对抗网络构建特征提取模型,学习目标的主要特征以及细微特征,再运用生成对抗网络生成多目标的运动轨迹,最终融和目标运动信息和外观信息,得到跟踪目标的最优匹配。在MOT16数据集下的实验结果表明,提出的多目标跟踪算法具有较高的精确度和鲁棒性,对比目前身份交换和跳变最少的算法,跳变的次数少了65%,准确度提高了0.25%。

关 键 词: 多目标跟踪;生成对抗网络;目标检测;路径预测;特征融合

中图分类号: TP 391

文献标志码: A

文章编号: 1005-3026(2020)12-1673-08

A Multi-target Pedestrian Tracking Algorithm Based on Generated Adversarial Network

WEI Ying¹, XU Chu-qiao¹, DIAO Zhao-fu¹, LI Bo-qun²

(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China; 2. School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China. Corresponding author: LI Bo-qun, E-mail: lbqhylyxab@163.com)

Abstract: In the field of multi-target tracking, the problems of target loss, identity exchange and switch are easy to occur under the conditions of complex background, target occlusion, target scale and attitude change. To solve these problems, a multi-target tracking algorithm was proposed based on detection. A human body and face association algorithm based on YOLO was used to classify and detect the position of the current frame objects, and the feature extraction model based on generative adversarial network was proposed to learn the main features and subtle features of the objects. Then the generative adversarial network was used to generate the motion trajectories of multiple targets, and finally the target's motion and appearance information were merged to obtain the optimal matching of the target tracking results. The experimental results show that the multi-target tracking algorithm proposed is both accurate and robust. Compared with the current algorithms with the least ID switch, the number of ID switch is 65% less and the accuracy is improved by 0.25%.

Key words: multi-target tracking; generative adversarial network; object detection; path prediction; feature fusion

在计算机视觉研究领域里,目标跟踪^[1]是主要的方向之一,有单目标跟踪和多目标跟踪两种类型。其中,多目标跟踪需要在给定的视频序列中同时标记数个目标,从而获得它们的运动轨迹。多目标跟踪在机器人导航、智能视频监控、自动驾驶等范围都有着极为普遍的运用。行人目标作为一

种典型的非刚体目标,跟踪难度较大,是实际应用中最常见的一种。

近年来,基于神经网络的深度学习技术取得极大的发展,具有代表性的检测算法包括 Fast R-CNN^[2], SSD^[3]和YOLO^[4]算法等。随着目标检测技术的进步,基于检测的多目标跟踪算法

(tracking-by-detection) 占据主要地位. 算法在每一帧中检测出目标, 然后与已有的跟踪轨迹进行匹配. 对于当前帧中的新目标, 需要形成新的轨迹; 对于离开当前帧中视野的目标, 需要终止目标的轨迹.

多目标跟踪场景比较复杂, 需要处理目标的光照、变形、遮挡等问题. 跟踪过程中背景与目标之间会发生相互交互, 因此应用高性能的检测算法在多目标跟踪中极为重要. 在跟踪任务中, 通常用卡尔曼滤波来进行跟踪目标的轨迹预测, 但目标发生姿态变化时不能达到很好的跟踪效果. 在跟踪目标与检测目标进行数据关联计算时, 一般通过匈牙利算法进行边界框重叠 (IOU)^[5] 的关联度量, 这种关联度量在状态估计不确定性高时, 容易出现身份交换和跳变的问题.

为了更好地应对上述多目标跟踪问题中的难题, 许多学者基于深度学习理论提出了不同措施, 以提高算法的性能. Wang 等^[6] 率先将深度学习应用到多目标跟踪中, 使用了自动编码器网络, 优化提取到的视觉特征, 并采用支持向量机来处理关联问题. Wojke 等^[7] 提出 Deep Sort 算法, 运用一个残差网络结构来提取目标的外观信息, 用匈牙利算法将外观特征向量的余弦距离与运动信息关联起来. Sadeghian 等^[8] 引入循环神经网络, 将 LSTM 提取的特征相融合, 获得相似度得分. 自从生成对抗网络模型^[9] 被首次提出以来, 文献[10] 运用生成对抗网络进行数据增强, 将其应用到行人重识别领域. 文献[11–12] 在有关预测行人运动轨迹的工作中, 通过结合生成对抗网络和 LSTM 来帮助提高预测效果.

针对上述观察, 本文提出了一个多目标跟踪算法的框架, 基于 YOLO 的人体人脸关联算法进行目标检测, 可以解决在密集场所中人体和人脸匹配困难问题, 提高行人目标检测的准确度; 在特

征提取模块和路径预测模块均引入了生成对抗网络, 对目标形状颜色等外观特征进行有效表达, 可以应对目标复杂多变的运动轨迹; 优化了跟踪与检测的数据关联算法, 在匹配时融合了外观信息和运动信息, 提高了整个模型的鲁棒性.

1 算法框架

本文提出的整体算法框架由 4 个模块组成, 分别是检测模块、特征提取模块、预测模块和匹配模块. 如图 1 所示, 首先对被跟踪视频序列的当前帧图像进行检测操作, 获取所有目标的位置信息, 即相互关联的人体检测框和人脸检测框, 人脸框的存在可以使人体框较为粗略的特征有所补充. 特征提取模块包含两种提取特征的网络, Net1 为基于生成对抗的行人特征提取网络, Net2 为常见的人脸识别网络, 两个特征拼接形成最终的特征. 同时使用基于生成对抗的行人多目标轨迹预测网络对每个目标的运动轨迹进行状态估计. 将以上信息送入最后的匹配模块, 进行轨迹更新, 以达到对每个目标的持续跟踪.

1.1 基于 YOLO 的人体人脸关联检测算法

本文提出了一种基于 YOLO 的人体人脸相关联的目标检测算法, 主要解决密集场所中行人目标检测困难问题. 在目标人体的外观相似时, 增加了人脸特征以增加外观特征的分度. 本文将 YOLO 的网络进行改进, 网络结构图如图 2 所示. 首先将检测图片送入网络中, 输出层包括 3 个不同尺度的特征图, 保证了模型对各种尺度物体的检测能力. 将包含特征的向量根据置信度进行降序排序, 先将 top 1 置信度的框的位置信息 (bounding box, 简称 bbox) 遍历其他 bbox 进行 IOU 计算. 如果值大于阈值, 则认为该 bbox 为重复框, 将其剔除. 然后再从剔除后剩余的 bbox 取

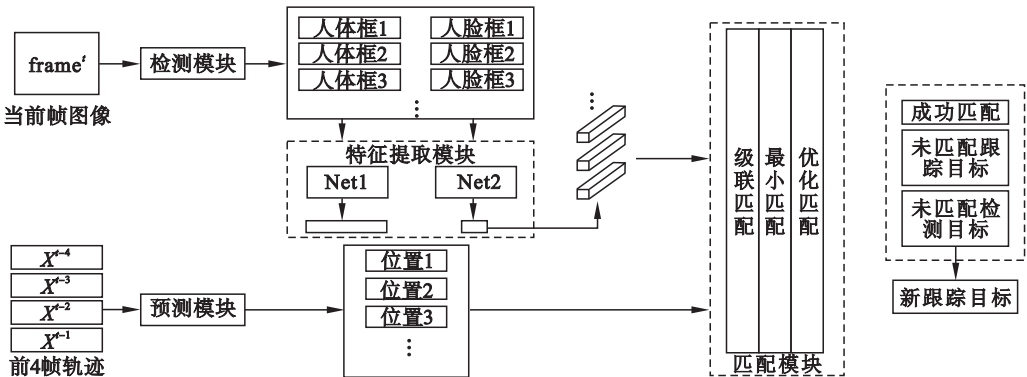


图 1 多目标跟踪算法架构

Fig. 1 Multi-target tracking algorithm architecture

出 top 2 的 bbox 重复以上的操作,直至遍历结束,最终得到精简的检测结果。

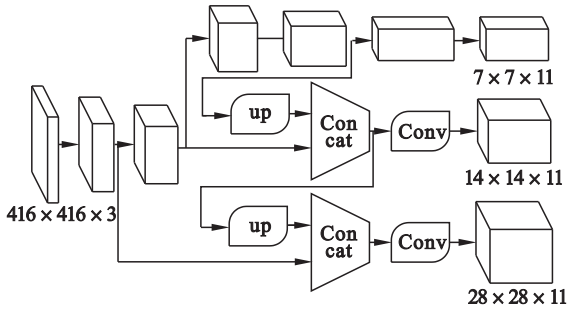


图 2 基于 YOLO 的人体人脸检测网络结构图

Fig. 2 Structure of YOLO-based human body and face detection network

改进后的输出层在原来的基础上增加了 4 维用于存放于人体框相关联的人脸框的位置信息,分别为相对人体框的人脸框的横向位置,纵向位置,宽度和高度信息。

$t_x^{\text{person}}, t_y^{\text{person}}, t_w^{\text{person}}, t_h^{\text{person}}$ 相当于输出特征的前 4 维,而 $t_x^{\text{person_face}}, t_y^{\text{person_face}}, t_w^{\text{person_face}}, t_h^{\text{person_face}}$ 相当于输出特征的后 4 维。当检测物体是人脸时,则不进行相关的计算。本文采用了更稳定的 L1 损失,损失函数如下:

$$\text{loss}_{x,y}^{\text{person}} = \lambda_{x,y}^{\text{person}} \sum \sum \left[|x_i^{\text{person}} - \hat{x}_i^{\text{person}}| + |y_i^{\text{person}} - \hat{y}_i^{\text{person}}| \right], \quad (1)$$

$$\text{loss}_{x,y}^{\text{face}} = \lambda_{x,y}^{\text{face}} \sum \sum \left[|x_i^{\text{face}} - \hat{x}_i^{\text{face}}| + |y_i^{\text{face}} - \hat{y}_i^{\text{face}}| \right], \quad (2)$$

$$\text{loss}_{w,h}^{\text{person}} = \gamma_{w,h}^{\text{person}} \sum \sum \left[\left| \sqrt{w_i^{\text{person}}} - \sqrt{\hat{w}_i^{\text{person}}} \right| + \left| \sqrt{h_i^{\text{person}}} - \sqrt{\hat{h}_i^{\text{person}}} \right| \right], \quad (3)$$

$$\text{loss}_{w,h}^{\text{face}} = \gamma_{w,h}^{\text{face}} \sum \sum \left[\left| \sqrt{w_i^{\text{face}}} - \sqrt{\hat{w}_i^{\text{face}}} \right| + \left| \sqrt{h_i^{\text{face}}} - \sqrt{\hat{h}_i^{\text{face}}} \right| \right], \quad (4)$$

$$\text{loss}_{x,y}^{\text{person_face}} = \lambda_{x,y}^{\text{person_face}} \sum \sum \left[|x_i^{\text{person_face}} - \hat{x}_i^{\text{person_face}}| + |y_i^{\text{person_face}} - \hat{y}_i^{\text{person_face}}| \right], \quad (5)$$

$$\text{loss}_{w,h}^{\text{person_face}} = \gamma_{w,h}^{\text{person_face}} \sum \sum \left[|w_i^{\text{person_face}} - \hat{w}_i^{\text{person_face}}| + |h_i^{\text{person_face}} - \hat{h}_i^{\text{person_face}}| \right]. \quad (6)$$

其中: $\text{loss}_{w,h}^{\text{person}}$ 是人体检测损失函数; x_i^{person} 是预测的行人相对横向位置; $\hat{x}_i^{\text{person}}$ 是对应的真实标签; y_i^{person} 是预测的行人相对纵向位置; $\hat{y}_i^{\text{person}}$ 为对应的真实标签。 $\text{loss}_{x,y}^{\text{face}}$ 为人脸检测损失函数; x_i^{face} 是算法预测的人脸相对横向位置; y_i^{face} 是人脸的相对纵向位置; $\hat{x}_i^{\text{face}}, \hat{y}_i^{\text{face}}$ 为真实标签。 $w_i^{\text{person}}, h_i^{\text{person}}$

是人体预测的相对宽度和高度。 $\hat{w}_i^{\text{person}}, \hat{h}_i^{\text{person}}$ 为对应的标签。 $w_i^{\text{face}}, h_i^{\text{face}}$ 是人脸预测的相对宽度和高度, $\hat{w}_i^{\text{person}}, \hat{h}_i^{\text{person}}$ 为对应的标签。 $\lambda_{x,y}^{\text{person}}, \lambda_{x,y}^{\text{face}}, \gamma_{w,h}^{\text{person}}, \gamma_{w,h}^{\text{face}}$ 为参数。

1.2 基于生成对抗的特征提取算法

在特征提取模块中,本文采用了基于生成对抗的算法提取行人特征。相比于一般的深度学习特征提取方法,通过生成对抗生成新的数据,使特征提取的网络在最大程度上减小相同 ID 图像间的类内特征变化和区分不同 ID 的图像间的类间特征。本文使用编码器作为识别学习的骨干网络,并利用不同条件下生成的图像,学习到目标的主要特征以及精细特征。

用 $X = \{x_i\}_{i=1}^N$ 和 $Y = \{y_i\}_{i=1}^N$ 表示真实的图像和其对应的标签, N 表示图像的数目, $y_i \in [1, K]$, K 为在数据集上所识别的 ID 数目,选取训练集中两幅真实图像 x_i 和 x_j 生成一幅新的图像,在生成模块交换它们的结构编码或外观编码。如图 3 所示,生成模块 $G(a_i, s_j) \rightarrow x_j'$ 由外观编码模型 $E_a: x_i \rightarrow a_i$ 以及结构编码模型 $E_s: x_j \rightarrow s_j$ 构成,结构编码使目标的几何和位置特征得到保留。并利用判别模型对后来生成的图像和原有的真实图像进行判别。

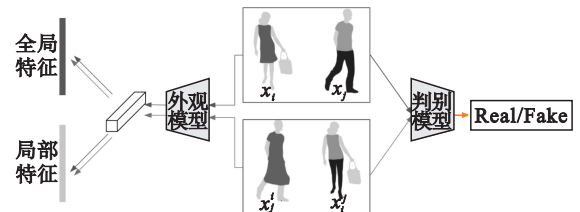


图 3 特征以及判别模型

Fig. 3 Features and discriminative model

对于不同 ID 的图像生成(图 4),给定两个图像 x_i 和 x_j ,所生成的图像 $x_j' = G(a_i, s_j)$ 需要分别保留来自 x_i 的外观编码 a_i 和 x_j 的结构编码 s_j 。然后,应能够在编码生成的图像后重建两个潜在编码,并根据图像的编码对生成的图像强制加上 ID 损失函数,以保持标识 ID 的一致性:

$$L_{\text{recon}}^{\text{code}_1} = E[\| a_i - E_a(G(a_i, s_j)) \|_1], \quad (7)$$

$$L_{\text{recon}}^{\text{code}_2} = E[\| s_j - E_s(G(a_i, s_j)) \|_1], \quad (8)$$

$$L_c^{\text{code}} = E[-\log(p(y_i | x_j'))]. \quad (9)$$

其中 $p(y_i | x_j')$ 是 x_j' 的预测概率,属于 x_i 的真实标签类 y_i ,该图像在生成 x_j' 时提供编码。此外,本文采用对抗性损失函数来匹配生成图像的分布与真实数据分布:

$$L_{\text{adv}} = E[\log D(x_i) + \log(1 - D(G(a_i, s_j)))]. \quad (10)$$

使用同一身份 ID 的任意两个图像之间进行图像的重构,如图 5 所示,以减少类内特征变化. 给定图像 x_i ,生成模块首先学习如何从自身重构 x_i . 此外,生成器应该能够通过具有相同标识 $y_i = y_i$ 的图像 x_i 来重构 x_i ,使用 ID 损失来区分不同的标识 ID:

$$L_{\text{recon}}^{\text{img1}} = E[\| a_i - E_a(G(a_i, s_j)) \|_1], \quad (11)$$

$$L_{\text{recon}}^{\text{img2}} = E[\| x_i - G_s(a_i, s_i) \|_1], \quad (12)$$

$$L_{\text{id}}^s = E[-\log(p(y_i | x_i))]. \quad (13)$$

其中 $p(y_i | x_i)$ 是图像外观编码属于真实标签类别的预测概率.

使用监督模型动态分配标签 x_j^i ,这取决于它从 x_i 和 x_j 得到的外观编码和结构编码. 判别模块方面,为了使其获得图像主要特征的识别能力,本文通过使其预测的概率分布 $p(x_j^i)$ 与监督预测的概率分布 $q(x_j^i)$ 之间的信息散度最小化,来对判别模块进行训练:

$$L_{\text{prim}} = E \left[- \sum_K q(k | x_j^i) \log \left(\frac{p(k | x_j^i)}{q(k | x_j^i)} \right) \right]. \quad (14)$$

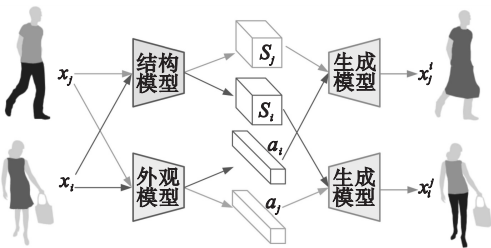


图 4 不同 ID 生成图像示意图

Fig. 4 Schematic diagram of different ID generation images

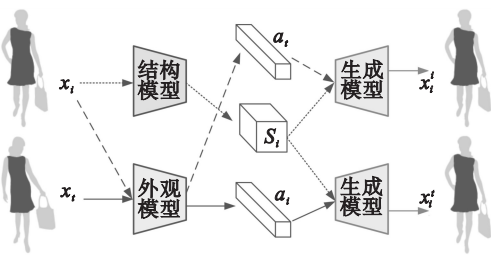


图 5 相同 ID 生成图像示意图

Fig. 5 Schematic diagram of the same ID generation image

本文提供另一种方法来替代生成分支,通过模拟图像中行人目标的服装变化,来代替使用生成的数据,进行主要特征的学习,当对以这种方式进行训练时,判别模块能够学习与服装无关的细微的 ID 相关属性. 把不同结构编码和外观编码组合生成的图像,视为提供结构编码的真实图像的同一种类. 对这个实现图像细微特征挖掘判别模块,使用标识 ID 损失进行训练:

$$L_{\text{fine}} = E[-\log(p(y_i | x_j^i))]. \quad (15)$$

为了优化总的目标,使用以下损失的加权和,对外观编码器、结构编码器、解码器和判别器共同训练:

$$L_{\text{total}}(E_a, E_s, G, D) = \lambda_{\text{img}} L_{\text{recon}}^{\text{img}} + L_{\text{recon}}^{\text{code}} + L_{\text{id}}^s + \lambda_{\text{id}} L_{\text{id}}^c + L_{\text{adv}} + \lambda_{\text{prim}} L_{\text{prim}} + \lambda_{\text{fine}} L_{\text{fine}}. \quad (16)$$

其中: $L_{\text{recon}}^{\text{img}} = L_{\text{recon}}^{\text{img1}} + L_{\text{recon}}^{\text{img2}}$ 是图像重建中的自我(同 ID)判别损失; $L_{\text{recon}}^{\text{code}} = L_{\text{recon}}^{\text{code1}} + L_{\text{recon}}^{\text{code2}}$ 是跨身份(不同 ID)生成中的编码重构损失; $\lambda_{\text{img}}, \lambda_{\text{id}}, \lambda_{\text{prim}}, \lambda_{\text{fine}}$ 是控制相关损失项重要性的权重.

使用数据集生成的图像如图 6 所示,其中第 1 行为原始图像,第 2 行为使用两个同一身份 ID 图像重构的图像,同时保留了目标的外观和结构特征. 其他为由两个不同 ID 的图像生成的图像,生成的图像出现服装配饰等方面的外观变化,保留目标自身的结构特征.



图 6 Market-1501 数据集生成图像示例

Fig. 6 Example of Market-1501 dataset generation image

1.3 基于生成对抗模型的多目标路径预测算法

多目标跟踪的实际场景中,行人多目标的轨迹预测时需要考虑运动的实际情况,周围人的活动也会影响目标的行走路径. 本文采用了基于生成对抗模型的多目标路径预测算法,应对复杂的人类交互,预测未来轨迹. 算法基于生成对抗的编码器-解码器结构,并提出一种池化模块来模拟行人之间的相互作用. 将目标与周围数个干扰目标的相对位置作为模块的输入,经过 MLP 和 Max - Pooling 处理,最终得到一个汇集了目标行人与周围行人位置信息的向量,以此模拟目标与周围人的交互.

本文的路径预测模型如图 7 所示,整体由 3 个主要部分构成:生成器、池化模块和判别器. 生成器基于编码以及解码的 LSTM 框架,采用池化模块对编码和解码的隐藏状态进行连接. 最后送入判别器进行判定轨迹是否为真.

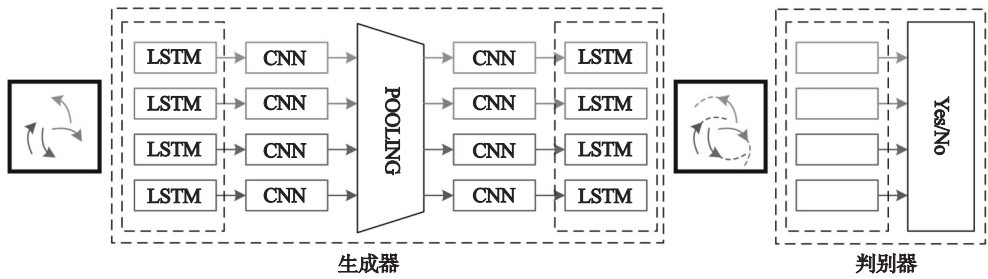


图 7 基于生成对抗多目标路径预测

Fig. 7 Multi-target path prediction model based on GAN

生成器部分,将每个目标的位置输入到作为编码器的 LSTM 单元,得到一个定长向量 e'_i ,引入以下循环:

$$e'_i = \phi(x'_i, y'_i; W_{ec}), \quad (17)$$

$$h'_{ei} = \text{LSTM}(h'_{ei-1}, e'_i; W_{\text{encoder}}), \quad (18)$$

其中: t 是序列; i 是目标; $\phi(\cdot)$ 是具备 ReLU 非线性的嵌入函数; W_{ec} 是嵌入权重; W_{encoder} 是 LSTM 的权重.

本文使用池化模块来模拟来往行人之间的交互作用,在可观测时刻之后,将场景中所有人的隐藏状态汇合起来,每个人获得一个合并的张量.通过初始化解码器的隐藏状态来调节输出轨迹的生成:

$$c'_i = \gamma(P_i, h'_{ei}; W_c), \quad (19)$$

$$h'_{di} = [c'_i, Z]. \quad (20)$$

其中: $\gamma(\cdot)$ 是包含 ReLU 非线性的多层感知器 (MLP); W_c 是嵌入权重,后续预测情况如下:

$$e'_i = \phi(x'^{t-1}_i, y'^{t-1}_i; W_{ed}), \quad (21)$$

$$P_i = \text{PM}(h'^{t-1}_{di}, \dots, h'^{t-1}_{dn}), \quad (22)$$

$$h'_{di} = \text{LSTM}(\gamma(P_i, h'_{di}), e'_i; W_{\text{decoder}}), \quad (23)$$

$$(\hat{x}'_i, \hat{y}'_i) = \gamma(h'_{di}). \quad (24)$$

其中: $\phi(\cdot)$ 是具备 ReLU 非线性的嵌入函数; W_{ed} 是嵌入权重.

判别器由一个解码器组成,输入为 $T_{\text{real}} = [X_i, Y_i]$, $T_{\text{fake}} = [X_i, \hat{Y}_i]$ 并将它们归类为真或假,在解码器的最后隐藏状态上运用多层感知机以得到最终的分分类分数.采用了一种多重损失函数,能够激励网络生成不同的样本.在 $N(0, 1)$ 中随机抽样 z 并使用 L_2 意义上的“最佳”预测作为本文的预测,生成 k 个候选的输出预测.

$$L_{\text{variety}} = \min_k \|Y_i - \hat{Y}_i^k\|_2. \quad (25)$$

1.4 匹配模块

本文中采用的匹配模块首先对目标运动信息进行匹配,具体的做法为计算轨迹预测模块的结果与检测结果之间的马氏距离:

$$d^{(1)}(i, j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i). \quad (26)$$

其中: d 为第 j 个检测结果的位置; y_i 为第 i 个跟踪器对跟踪目标的预测位置; S_i 为检测位置 and 平均跟踪位置的协方差矩阵.设定阈值 $t^{(1)}$,当此次关联的马氏距离小于它时,运动状态关联成功,关联度量为

$$b_{i,j}^{(1)} = 1[d^{(1)}(i, j) \leq t^{(1)}]. \quad (27)$$

在运动不确定度较高时,如长时间跟踪或出现长时间遮挡的情况,引入外观特征进行匹配.外观特征即人体框人脸框的联合特征.通过将每一个跟踪的目标的历史特征构造成一个特征库,存储最近成功关联的帧的特征,计算待匹配的特征与特征库特征之间的余弦距离最小值进行匹配:

$$d^{(2)}(i, j) = \min\{1 - r_j^T r_k^{(i)} | r_k^{(i)} \in R_i\}. \quad (28)$$

如果最小距离小于设定阈值,则关联成功.使用两种度量的方式进行加权得到.运用组合距离阈值判断不等式,作为判断第 i 个目标跟踪结果和第 j 个目标检测结果之间是否关联的总公式:

$$c(i, j) = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j), \quad (29)$$

$$b_{i,j} = \prod_{m=1}^2 b_{i,j}^{(m)}. \quad (30)$$

可以看出,只有当 $c(i, j)$ 同时满足两个度量的阈值的要求,才设定为完成了正确的关联.马氏距离度量对短时跟踪效果较好,外观特征度量对长时跟踪或长时间遮挡的情况更有效.可以针对不同的任务设定不同的 λ 进行适应.

2 实 验

2.1 与当前主流算法进行比较

本文使用 MOT16^[13] 基准测试数据集评估了所提出的跟踪算法的性能,并与 Deep Sort^[7]、Sort^[14] 等先进算法进行了各项指标的对比. MOT16 数据集具备多种多样的数据类型,具有在不同的视线角度、相机运动方式以及不同天气状况下拍摄的画面.根据 MOT16 的评估标准,实验结果如表 1 所示,多目标跟踪准确度 (MOTA) 指标位于第 2 名,多目标跟踪精确度 (MOTP) 最高,

比第 2 名高了 0.25%,比同类的基于深度学习的 Deep Sort 提高了 1.64%. 准确度(MOTA)与身份跳变数目(IDS)对比如图 8 所示,在不影响跟

踪准确度的条件下,本文提出的算法身份交换和跳变明显少于其他算法. 如图 9 所示,虚警数(FP)、漏警数(FN)明显降低.

表 1 多目标跟踪算法跟踪结果
Table 1 Multi-target tracking algorithm tracking results

算法	MOTA	MOTP	IDS	FP	FN
DNT	68.2	79.4	933	11 479	45 605
LMP_p	71.0	80.2	434	7 880	44 564
MCMOT_HDM	62.4	78.3	1 394	9 855	57 257
NOMTwSDP16	62.2	79.6	406	5 119	63 352
EAMTT	52.5	78.8	910	4 407	81 223
POI	66.1	79.5	805	5 061	55 914
Sort	59.8	79.6	1 423	8 698	63 245
Deep Sort	61.4	79.1	781	12 852	56 668
本文	65.7	80.4	152	3 251	42 051

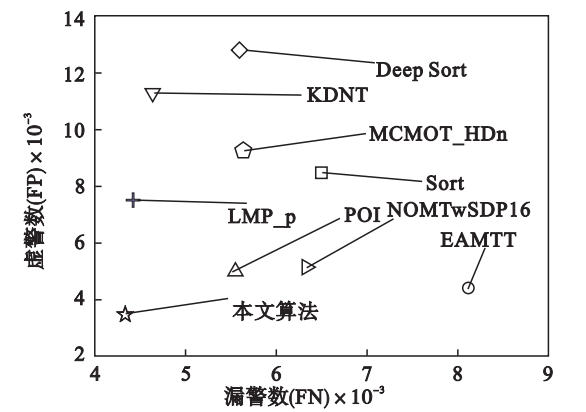
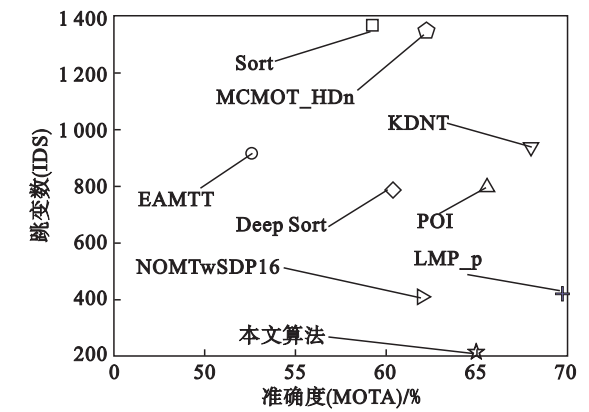


图 8 准确度与身份跳变数对比
Fig. 8 Comparison of accuracy and identity jump

图 9 漏警数与虚警数对比
Fig. 9 Comparison of FN and FP

2.2 实施细节

本文算法框架中的目标检测模块、特征提取模块、路径预测模块在目标检测数据集 ImageNet、行人重识别数据集 Market - 1501、行人视频数据集 Eth 中分别进行训练,得到最优的权重后再融入到整体的框架中. 目标检测模块中 $loss_{w_{f_i}}$ 采用均方差误差,其他的损失为交叉熵损失,并采用 $L1$ 正则化.

在特征提取模块的整个训练过程中固定权重 $\lambda_{img}=5, \lambda_{id}=0.5$. 用作区分特征学习损失 L_{prim} 和 L_{fine} ,直到生成器稳定下来. 本文模型在 Market - 1501 上进行 30 000 次迭代后,再将两个损失求和,随后的 4 000 次迭代中把 λ_{prim} 从 0 线性增加到 2,并设定 $\lambda_{fine}=0.2\lambda_{prim}$.

在跨身份(不同 ID)图像生成中,在生成图像之前训练 E_a, E_s 和 G ,在生成图像之后训练 E_a, E_s 和 D . 实验发现匹配模块中马氏距离匹配阈值取 9.487 7 最佳. 外观特征采用人体框人脸框的联合

特征,其中人体框特征包含人体的主要特征和细微特征,人脸特征作为补充进行融合. 将每一个跟踪目标的历史特征构造成一个特征库,将最近 100 个成功关联的帧的特征进行存储,计算待匹配的特征与特征库特征之间的余弦距离最小值进行匹配.

2.3 消融实验

为进一步分析所提方法各部分的有效性,在基于普通人体特征和卡尔曼滤波跟踪器的基础上,设计了消融实验来对算法框架中的各个部分进行对比分析,结果如表 2 所示.

通过对比准确度与身份跳变指标,在检测跟踪过程中增加了人脸特征之后,跟踪器的准确度有所提升,身份交换和跳变情况有了明显的缓解. 进一步应用通过生成对抗网络提取的增强人体特征代替普通人体特征,多目标跟踪的准确度基本不变,但是身份交换和跳变数目降低了 23%. 最后再用基于生成对抗网络的路径预测模块代替传统

的卡尔曼滤波算法,可以看出本文改进的算法在身份交换和跳变数目上进一步降低 26%,达到最低。

表 2 消融实验结果
Table 2 Ablation experiments results

算法及指标	MOTA	MOTP	IDS	FP	FN
普通人体特征 + 卡尔曼路径预测	61.4	79.1	781	12 852	56 668
普通人体特征 + 人脸特征 + 卡尔曼路径预测	62.3	80.2	422	8 542	51 453
增强人体特征 + 人脸特征 + 卡尔曼路径预测	62.3	79.5	324	5 442	50 786
增强人体特征 + 人脸特征 + 生成对抗路径预测	62.3	79.8	241	4 385	44 512

本文算法通过增加人脸特征,提高了检测的准确性;通过引入主要特征和细微特征结合的增强人体特征,增强了图像特征的表现力;应用基于生成对抗网络的路径预测算法生成目标轨迹,得到目标更准确的位置序列.有效解决了现存算法中,检测结果与跟踪路径不匹配,身份变换频繁的问题.

2.4 定性分析

图 10 为本文算法在 MOT 数据集中一段视频序列上跟踪的实验结果.图10所示的序列



图 10 MOT 序列跟踪结果
Fig. 10 MOT sequence tracking results

中,行人目标背景较为复杂,目标数量较多,目标间存在着频繁的交互.目标运动过程中发生了由远及近和由近及远的变化,使目标尺度发生改变.目标还出现了遮挡现象,以及随后消失又重现的情况.如图 10 所示,本文取得了良好的跟踪效果.在背景复杂、目标遮挡、尺度姿态变化的应用场景中,有极大的抗干扰能力,有效解决了跟踪偏移和匹配错误的问题,实现目标平稳跟踪.

3 结 论

本文针对多目标跟踪中背景复杂、目标遮挡、目标尺度和姿态变化情况下,容易出现目标丢失、身份交换和跳变的问题,提出了一种基于生成对抗网络的多目标跟踪算法.通过使用基于 YOLO 的人体人脸关联算法,对当前帧待检目标进行检测,提出了基于生成对抗网络的特征提取模型,且引入了人脸特征,使对目标的特征表示更加鲁棒.再使用生成对抗网络生成复杂交互下更准确的多目标的运动轨迹,在匹配模块中结合目标的运动信息和外观信息,得到最终的目标跟踪结果.实验结果表明,在出现背景复杂、目标遮挡、尺度变化等干扰情况时,本文算法都能平稳且准确地对目标进行跟踪,且大幅度减少了目标身份跳变情况的发生,具有较高的精确度.

参考文献：

[1] 李玺,查宇飞,张天柱,等. 深度学习的目标跟踪算法综述 [J]. 中国图象图形学报,2019,24(12):2057-2080.
(Li Xi, Zha Yu-fei, Zhang Tian-zhu, et al. Survey of visual object tracking algorithms based on deep learning[J]. *Journal of Image and Graphics*,2019,24(12):2057-2080.)

[2] Ren S Q,He K M,Girshick R,et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]//Advances in Neural Information Processing Systems. Montreal,2015:91-99.

[3] Wei L, Dragomir A, Dumitru E, et al. SSD: Single shot multibox detector [C]//European Conference on Computer Vision. Amsterdam,2016:21-37.

[4] Redmon J, Farhadi A. Yolo9000: better, faster, stronger [C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,2017:7263-7271.

[5] Bochinski E,Eiselein V, Sikora T. High-speed tracking-by-detection without using image information[C]//International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017. Lecce,2017:1-6.

[6] Wang L,Pham N T,Ng T T,et al. Learning deep features for multiple object tracking by using a multi-task learning strategy [C]//IEEE International Conference on Image Processing. Paris,2014:838-842.