Opto-Electronic Engineering
光 电 工 程

Article

2020 年，第 47 卷，第 1 期

DOI: 10.12086/oee.2020.190136

基于 R-FCN 框架的多候选关联在线多目标跟踪

鄂　贵，王永雄*

上海理工大学光电信息与计算机工程学院，上海 200093

摘要：在线多目标跟踪是实时视频序列分析的重要前提。针对在线多目标跟踪中目标检测可靠性低、跟踪丢失较多、轨迹不平滑等问题，提出了基于 R-FCN 网络框架的多候选关联的在线多目标跟踪模型。首先，通过基于 R-FCN 网络从 KF 预测结果和检测结果中获取更可靠的候选框，然后利用 Siamese 网络进行基于外观特征的相似性度量，实现候选与轨迹之间的数据关联，最后通过 RANSAC 算法优化跟踪轨迹。在人流密集和目标被部分遮挡的复杂场景中，提出的算法具有较高的目标识别和跟踪能力，大幅减少漏检和误检现象，跟踪轨迹更加连续平滑。实验结果表明，在同等条件下，与当前已有的方法对比，本文提出在目标跟踪准确度(MOTA)、丢失轨迹数(ML)和误报次数(FN)等多个性能指标均有较大提升。

关键词：多目标跟踪；候选模型；孪生网络；轨迹估计

中图分类号：TP391　　**文献标志码：**A

引用格式：鄂贵，王永雄. 基于 R-FCN 框架的多候选关联在线多目标跟踪[J]. 光电工程，2020，**47**(1): 190136

Multi-candidate association online multi-target tracking based on R-FCN framework

E Gui, Wang Yongxiong*

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: Online multi-target tracking is an important prerequisite for real-time video sequence analysis. Because of low reliability in target detection, high tracking loss rate and unsmooth trajectory in online multi-target tracking, an online multi-target tracking model based on R-FCN (region based fully convolutional networks) network framework is proposed. Firstly, the target evaluation function based on R-FCN network framework is used to select more reliable candidates in the next frame between KF and detection results. Second, the Siamese network is used to perform similarity measurement based on appearance features to complete the match between candidates and tracks. Finally, the tracking trajectory is optimized by the RANSAC (random sample consensus) algorithm. In crowded and partially occluded complex scenes, the proposed algorithm has higher target recognition ability, greatly reduces the phenomenon of missed detection and false detection, and the tracking track is more continuous and smooth. The experimental results show that under the same conditions, compared with the existing methods, the performance indicators of the proposed method, such as target tracking accuracy (MOTA), number of lost trajectories (ML) and

收稿日期：2019-03-25；　收到修改稿日期：2019-06-27

基金项目：国家自然科学基金资助项目(61673276，61703277)

作者简介：鄂贵(1994-)，男，硕士研究生，主要从事目标跟踪的研究。E-mail：332437798@qq.com

通信作者：王永雄(1970-)，男，博士，教授，主要从事智能机器人及视觉的研究。E-mail：wyxiong@usst.edu.cn

number of false positives (FN), have been greatly improved.

Keywords: multi-target tracking; candidate model; Siamese network; trajectory estimation

Citation: E G, Wang Y X. Multi-candidate association online multi-target tracking based on R-FCN framework[J]. *Opto-Electronic Engineering*, 2020, **47**(1): 190136

1 引 言

多目标跟踪是计算机视觉领域研究的热点之一，在视频分析、人机交互、自动驾驶和交通管制等众多领域有着十分广泛的应用。近年来，随着目标检测算法的快速发展和目标检测精度不断改善[1-2]，基于检测的多目标跟踪框架[3-9]得到了广泛关注和研究。基于检测的多目标跟踪框架流程是：首先使用离线训练好的目标检测器逐帧检测目标，利用相似性匹配方法对检测目标进行关联，然后不断地利用生成的轨迹与检测结果进行匹配，生成更加可靠的轨迹。

由于目标的多样化、运动方式的不确定性以及遮挡等原因，多个目标的运动轨迹关联比较复杂，因此，多目标检测与运动轨迹之间的关联方式是多目标跟踪的研究重点之一。为了处理多目标之间数据关联的难题，目前提出了多种应对措施，主要分为相邻的数据关联方法和多帧的数据关联方法两大类。由于相邻帧的数据关联方法采用在线学习的方式，在实际应用中更具有价值，因此被大量研究。Bewley[4]提出利用 Kalman 滤波器构建关于目标运动的线性预测模型，预测目标在下一帧的运动信息，然后通过匈牙利算法实现多目标在相邻帧之间的关联，此方法在一定程度上提高了目标与轨迹关联的成功率，但在比较复杂的场景中仅考虑了目标运动信息，算法的跟踪性能较差。考虑目标频繁被遮挡和外观相似等因素，Wojke[5]在 Bewley 的基础上做了改进，使用卷积神经网络的方法提取目标的高度抽象特征作为其外观信息，同时在数据关联中加入级联匹配策略，对目标的跟踪表现得到较好的提升，但对目标的漏检和误检未做分析。文献[6]在检测框架中建立关于判别目标漏检的模型，使漏检目标有机会重新被检测，有效地改善了目标漏检问题，但算法的计算复杂度明显增加。在线学习的方法采用不同方式处理了相邻帧之间的目标数据关联问题，只利用了单帧中的目标数据实现与运动轨迹之间的数据关联。但是，当目标检测结果不可靠时，若只考虑检测结果作为目标的候选，直接把下一帧的检测结果作为轨迹关联的对象，往往很难得到满意的跟踪结果。

为了解决检测的不可靠性问题，许多文献提出了多帧的数据关联方法[10-13]，此类方法是利用多帧检测的响应数据建立模型，可有效地抑制轨迹漂移和错误匹配等问题，主要有网络流模型、条件随机场模型、广义关联图模型等。Milan 等[10]提出一种离散-连续能量最小化的方法，将离散数据关联问题与轨迹估计的连续问题整合到一个能量函数中，用能量最小化算法求得最优解得到目标轨迹，但对运动相似且距离相近的目标区分度不好。文献[12]在 Dehghan 等[11]的基础上将关联问题转化为广义最大团问题，通过不断迭代寻找能量最小的子图解决关联问题，并考虑了轨迹片段长度的自适应，取得了较好的跟踪结果。虽然多帧数据关联方法能够获得不错的跟踪效果，但它属于离线的跟踪方法，且计算量偏大，不适合实时性要求高的目标跟踪应用。因此，为了获得跟踪的可靠性和实时性，不仅要求目标检测方法可靠，目标之间数据关联准确，而且要求部分目标遮挡时运动估计准确。

综上，本文提出了一种基于 R-FCN 框架的多候选关联的在线多目标跟踪方法。其创新之处在于，首先，提出了融合卡尔曼滤波器的预测框和检测器的检测框的候选选择模型，不再仅依靠检测结果作为候选框，提高了算法的鲁棒性。然后，利用 Siamese 网络框架[14]实现基于目标外观相似性度量，并融合了目标多种特征信息完成多目标之间数据关联，提高了对复杂跟踪场景中目标的判别能力。最后，针对人流密集以及容易被遮挡的复杂场景中目标可能存在的漏检和误检等问题，提出使用 RANSAC 算法对跟踪轨迹优化处理，从而得到更加完整准确的轨迹信息。

2 基于 R-FCN 的多目标跟踪框架

本文构建的基于检测的多目标跟踪框架主要分为候选框的选择、数据关联和轨迹优化等几个步骤，提出的算法流程如图 1 所示。

结合图 2 进一步说明算法流程。首先获取当前帧中的候选框，每个目标的候选框有两种可能的候选位置，分别来自检测器的检测(浅蓝色的实线框)和卡尔曼滤波器的预测(深红色的虚线框)。其次，通过基于

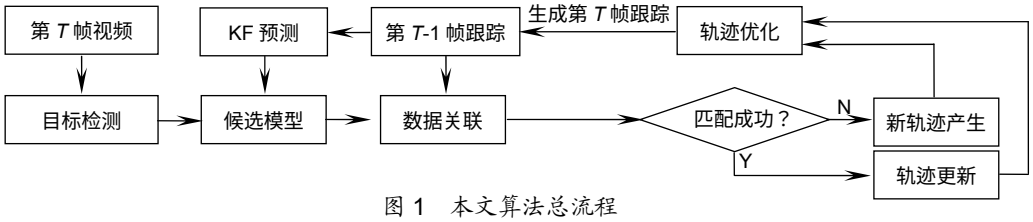


图 1 本文算法总流程

Fig. 1 The general flow of the algorithm

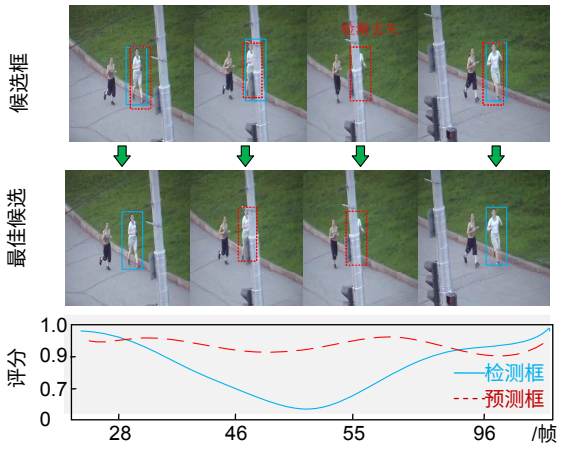


图 2 候选框选择流程图

Fig.2 Candidates selection flow chart

R-FCN 网络构建的目标评价函数对这些候选框进行评分(浅蓝色曲线和深红色曲线分别对应目标在不同帧中检测框得分和预测框得分)，获得判别这两种不同候选框是否足够可靠的依据，然后使用非极大值抑制和最低阈值的方法过滤一些可靠性较低的候选框，得到最终的目标候选框。接着，根据各个目标特征的重要性，把目标的空间位置信息、外观信息和目标尺度等多个特征的融合结果作为轨迹与最佳的候选框之间关联的依据。其中，外观特征通过 Siamese 网络框架获得。最后，为进一步改善目标漏检和误检问题，使用 RANSAC 算法对已有的轨迹进行优化。

2.1 建立目标候选选择模型

2.1.1 基于 R-FCN 的目标识别

R-FCN 是在 Faster RCNN 基础上改进的，在保证检测精度不变的情况下，极大地提高了检测速度，为了解决在检测网络中加入 ROI pooling 后，所生成的特征映射图对目标位置不再敏感的问题而提出的。R-FCN 结构如图 3 所示，主要由全卷积网络 FCN、区域生成网络 RPN 和 ROI 子网络三个部分构成。为了保证分类子网络和定位子网络所处理的目标特征映射图对目标位置比较敏感，提出了将目标的各部分位置信息融入 ROI pooling 的位置敏感的区域池化(position-sensitive ROI pooling)。R-FCN 实现目标检测的过程主要分为两步，首先利用去掉全连接层的 ResNet-101 网络处理待检测图片，生成关于目标位置不敏感的特征映射图，然后在生成的特征映射图上通过 RPN 获得不同目标在不同尺度下的搜索框，再利用非极大值抑制方法过滤得分较低的搜索框，完成目标的初步定位任务；其次，分别利用目标的分类子网络和定位子网络处理由 RPN 输出的目标位置信息，从而实现目标的准确分类和精确定位。

为了得到最佳的候选框，本文利用 R-FCN[15]网络框架对同一帧中获得的检测框进行重新识别，并通过 R-FCN 的输出结果构建检测框的目标函数。采用轻量级网络 SqueezeNet[16]作为 R-FCN 提取特征的基本骨架，减少生成目标特征映射图的时间，从而保证算法处理每帧中检测目标的实时性。把每个候选框作为一个感兴趣的区域，采用 $\boldsymbol{Z}=(x_0,y_0,w,h)$表示感兴趣区域，其中$(x_0,y_0)$表示区域左上角点坐标，$w$、$h$ 分别表示宽和高。利用一个 $k\times k$ 的网格将每个感兴趣区域划分为 k^2 部分，即 k^2 个 bin，每个 bin 对应目标空间位置的一部分。目标感兴趣区域 $\boldsymbol{Z}$ 的得分定义如下：

$$S_{\rm d}(c\,|\,r,z)=\boldsymbol{\varphi}\left(\sum_{i,j}^{k}\sum_{(x,y)\in bin_{(i,j)}}\frac{r_{i,j,c}(x,y)}{n_{i,j}}\right),\qquad(1)$$

其中：$\varphi(x)=\dfrac{1}{1+{\rm e}^{-x}}$，$r$ 为感兴趣区域的某一部分对应的特征映射图，用 (i,j) 表示感兴趣区域的第几部分，$n_{(i,j)}$ 为这部分像素点的总数，c 表示目标类别为行人。

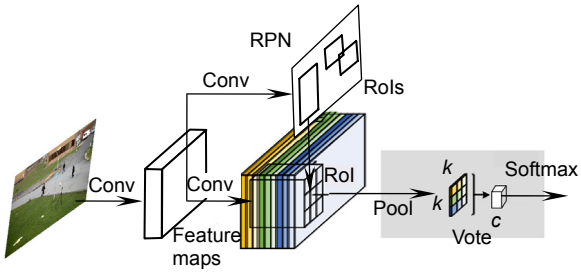


图 3 R-FCN 网络架构

Fig. 3 R-FCN network architecture

在训练阶段，训练的正样本从目标周围随机采样得到，相同数量的负样本以同样的方式从背景周围采样得到。通过将目标 k^2 个部分的空间位置信息融入目标类别得分中，最终得到健壮的目标类别得分。比如，k 取 3，分别表示目标的上左部分、上中部分、上右部分等 9 个部分。因此，当目标被部分遮挡时，通过 R-FCN 仍然能很好地识别目标。

2.1.2 构建候选框的评价函数

在多目标跟踪过程中，若目标发生遮挡，通过检测器难以得到准确的目标。当目标短时间被遮挡时，如图 2 中序列第 55 帧中目标被路标遮挡，此时目标检测丢失，但利用 KF 仍然可以获得不错的目标位置预测结果，可以通过 KF 的预测来抑制跟踪目标漂移。然而，当目标长时间被遮挡时，滤波器的预测精度会随着遮挡时间的延长而降低，因此，需要对滤波器预测结果进行可靠性评估。使用 $L_{\rm det}$ 表示某条轨迹的连续跟踪的检测数量，$L_{\rm trk}$ 表示某条轨迹自上次成功匹配后连续丢失的检测数量，一条完整轨迹可以由 $L_{\rm det}$ 和 $L_{\rm trk}$ 共同表示。轨迹评价函数定义如下：

$$S_{\rm trk}=1-\log(1+\gamma L_{\rm trk}) \quad 。 \tag{2}$$

为了对候选框有统一评价标准，提出了一种融合两种评价指标(式(1)和式(2))的评价函数。评价函数定义如下：

$$S=S_{\rm d}(c\,|\,r,z)T(z\in V_{\rm det})+S_{\rm trk}T(z\in V_{\rm trk}) \quad , \tag{3}$$

$$T(x)=\begin{cases}1, & x=\text{True}\\ 0, & x=\text{False}\end{cases} \quad , \tag{4}$$

其中：z 表示感兴趣区域，$V_{\rm det}$、$V_{\rm trk}$ 分别表示同一帧中收集的检测框集合和预测框集合。

2.2 基于 Siamese 网络提取的外观特征

在数据关联过程中，采用目标的外观特征进行匹配是一种重要的匹配度量方式。外观特征选取的优劣直接影响轨迹与检测框之间的分配结果，因此选取健壮的行人的外观特征尤为重要。为了比较两个目标外观之间的相似性，如图 4 所示，本文利用 Siamese 网络框架实现轨迹与候选框之间的外观相似度量。孪生网络框架由两个相同的网络分支和决策网络组成，其中每个网络分支用于获取对应输入的深度信息，而决策网络将两个分支获取的两个目标的深度信息融合得到两个目标外观相似度。

为了得到比较好的判别效果，同时保证跟踪的时效性，预训练模型选用深度残差网络 ResNet-50[18]，去除 ResNet-50 的平均池化层和全连接层，通过卷积层提取目标外观特征信息，并将 Spp network[17]加到网络的全连接之前适应不同尺寸图像的输入，减少输入图像的信息损失，从而得到更加健壮的特征信息。训练数据采用行人重新识别数据集 Market1501[19]，该数据集包含在不同视角和不同场景下拍摄的 1501 个不同的行人目标，总共 32000 张训练图片。为了保证网络模型更容易收敛，训练输入采用批量的方式进行处理，每次输入一批图片(32 张)对网络进行训练。为了更好地区分正负样本，采用对比损失作为 Siamese 网络在训练过程中的损失函数，每次从一批训练样本中挑选最难训练的一组样本进行训练，使正样本之间的欧氏距离尽量小，负样本之间的欧氏距离尽量大。

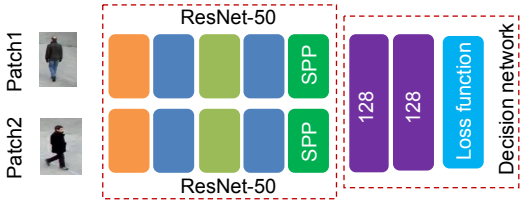


图 4 Siamese 网络结构

Fig. 4 Siamese network structure diagram

2.3 数据关联

考虑到多目标之间相互遮挡、目标形变以及目标外观存在突变等因素，在计算匹配代价时，我们融合了目标外观信息、目标尺寸信息以及运动信息，因此可以得到更为合理的匹配效果。目标外观特征由 2.2 节的 Siamese 网络得到，通过目标在下一帧中的 KF 预测框与候选框之间的交并比(IOU)计算目标的运动信息，目标尺寸信息由目标框的长和宽确定。

运动信息匹配代价：

$$A_{\rm m}(T_j,D_i^t)=\text{IOU}(T_j^{t'},D_i^t) \quad , \tag{5}$$

$$\text{IOU}(a,b)=\frac{a\cap b}{a\cup b} \quad , \tag{6}$$

其中：T_j 表示第 j 条轨迹，$T_j^{t'}$ 表示轨迹 T_j 在第 t 帧中的预测框位置，D_i^t 表示第 t 帧中第 i 个候选框，最低重叠率 σ 取 0.3。

外观特征匹配代价：

$$A_{\rm a}(T_j,D_i^t)=\frac{1}{t_1}\sum_{n=1}^{t_1}\boldsymbol{f}(T_j^n,D_i^t),\ t_1\in\{t_0,t-1\} \quad , \tag{7}$$

其中：$\boldsymbol{f}(T_j^n,D_i^t)$ 表示第 j 条轨迹在第 n 个位置的目标框与第 t 帧中第 i 个目标检测框之间的外观相似度，可以通过 Siamese 网络得到。

目标框尺度特征匹配代价：

$$A_s(T_j^{t'},D_i^t)=\exp\left(-\lambda\left\|\frac{w_2-w_1+h_2-h_1}{w_1+w_2+h_1+h_2}\right\|\right) \quad (8)$$

其中：h_1、w_1 分别表示 $T_j^{t'}$ 的长和宽，h_2、w_2 分别表示 D_i^t 的长和宽，λ 取 1.4。

因此，最终的多特征融合代价为

$$A(T_j,D_i^t)=\alpha A_m(T_j^{t'},D_i^t)+\beta A_a(T_j^{t'},D_i^t)+(1-\alpha-\beta)A_s(T_j^{t'},D_i^t) \quad (9)$$

其中：α、β 分别表示运动信息和外观信息所占总匹配代价的比例。

2.4 基于 RANSAC 算法的目标漏检处理

在某些帧中，由于目标可能被遮挡或者漏检，利用目标检测器和 KF 预测器都无法得到合适的候选框，因此通过轨迹与候选框之间的数据关联无法找到对应目标，此时目标的运动轨迹便不再是连续的。为了解决或部分解决上述问题，在目标数据关联之后，每隔 N 帧对由跟踪目标框的质心连接而成的轨迹进行优化。首先，收集每条待优化轨迹上所有点，利用 RANSAC 方法估计一条最佳轨迹，然后在同一条轨迹的不同位置添加一个虚拟节点，改善跟踪过程中存在的漏检和预测失败问题。当某条轨迹在某个位置匹配不到检测目标时，可以使用虚拟节点来代替丢失的检测目标，如图 5 所示。

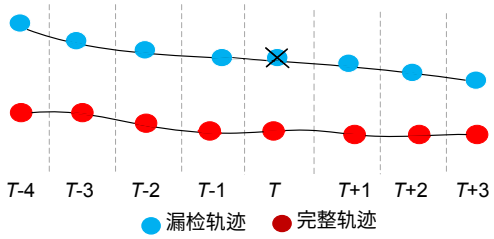


图 5 目标轨迹存在漏检

Fig. 5 Missing detection of target trajectory

用 V_s 表示一条轨迹上所有的节点，包含局内点和局外点。优化轨迹的关键是正确判别轨迹上的局内点和局外点。在短时间内，假设跟踪目标做直线运动。一条轨迹在某帧中的空间位置，可以通过如下方式建模：$\hat{P}(t)=a_1t+a_0$，a_0、a_1 均为二维向量，$P_s(t)$ 表示轨迹上的第 t 个跟踪位置。通过如下式子判断局内点和局外点：

$$V_s(\text{inliers})=\{V_s(t):|a_1t+a_0-P_s(t)|<\delta\} \quad (10)$$

其中δ表示阈值，取值为 5。为了使轨迹上局内点数量最大，通过如下公式优化：

$$(\hat{a}_0、\hat{a}_1)=\arg\max(V_s(\text{inliers})) \quad (11)$$

其中：$\hat{a}_0$、$\hat{a}_1$ 为最优参数。由于轨迹点由局内点和局外点两部分组成，可以利用 RANSAC 方法寻找一条包含更多局内点的最佳轨迹。当某条轨迹在第 t 个位置对应的检测丢失时，可以通过 $Q^t=a_1t+a_0$ 计算轨迹在第 t 个虚拟点位置信息，并修正丢失的目标框信息，从而得到更加完整的目标运动轨迹。

3 实验结果与分析

为了验证算法的有效性，在经典的目标跟踪数据集 MOT16 上进行实验，因该数据集提供相关视频序列的目标检测结果，在实验中不针对目标进行检测。MOT16 数据集囊括了 7 个训练集视频序列和 7 个测试集视频序列，这些视频序列包含运动相机、目标频繁遮挡、人流密集、多视角等挑战因素，属于比较复杂的跟踪场景。

实验采用基于 Linux 的硬件平台，以 Pycharm2018 为软件平台，工作站配置为 2.6 GHz CPU，32 G 内存，显卡为 GTX1080TiGPU。采用随机梯度下降法训练 R-FCN 网络参数，学习率为 0.001，在 MSCOCO 数据集上训练，输入的图像大小为 1152×640，每次批量输入 32 张图片，训练的最大迭代次数为 40k，目标位置敏感池化层部分 k 取 7，轨迹置信度参数γ为 1.4，非极大值抑制参数 T_{nms} 为 0.4，行人分类得分阈值 T_s 为 0.4，特征信息融合比例因子α、β 分别取 0.6 和 0.3。

3.1 定性分析

首先在 MOT16 测试集中的 7 个视频序列上进行实验，选择其中 3 个序列(MOT16-01、03 和 06)进行定性分析。这些视频序列是在不同场景不同光照条件下拍摄的，其中有些序列跟踪场景比较复杂，人流密集，相互遮挡严重，要实现目标的稳定准确跟踪难度较大。

MOT16-01 序列是在比较暗的场景下拍摄，人流适中，目标间交互过程中存在短时间遮挡，图 6(a)显示了 106#、264#、349#三帧视觉跟踪效果图，编号为 4 号、6 号的目标从 106#运动到 349#过程中存在短暂的漏检，但这两个目标仍然能持续准确地被跟踪。

MOT16-03 序列共有 1500 帧，分辨率比较高，帧数较快，是在灯光较亮的场景下拍摄的，跟踪场景比较复杂，人流较为密集，除少数漏检目标跟踪失败外，大部分目标都能被正常跟踪，说明本文算法对复杂环境有较好的鲁棒性，图 6(b)包含了 319#、424#两帧跟踪结果图，大部分目标(如 159 号、144 号、99 号、116

号、131 号等)从 319 帧运动到 424 帧都能被跟踪，而少数目标如 86 号、114 号、142 号等因长时间遮挡导致目标一直被漏检，同时其 KF 预测结果也未能及时更新，从而使得目标跟踪失败。137 号目标因被路灯遮挡与旁边外观相似的目标发生 IDS 转换。

MOT16-06 序列共有 1194 帧，是在人流较为密集的马路边上拍摄的。由于拍摄过程中相机是不断向前运动的，并且路边行人比较多，因此有很多目标频繁进出跟踪场景。图 6(c)中包含了其中的 473#、531#、1074#三帧跟踪结果图，对比 473#和 531#两帧图片，容易看到除了 336 号和 294 号目标因自身大部分遮挡和外观变化原因跟踪失败外，其他目标基本都能正常跟踪。观察 1074#，发现仅 283 号目标在之前两帧均有出现，并且目标 IDS 一直未发生变化，目标被持续稳定跟踪。

3.2 定量分析

3.2.1 评估标准

本文使用 CLEAR MOT[20]标准对算法定量评估，选取其中 5 个重要性能指标：多目标跟踪准确度(multiple object tracking accuracy，MOTA，用 A_{MOT} 表示)、目标被误报次数(false positive，FP，用 F_{P} 表示)、目标丢失次数(false negative，FN，用 F_{N} 表示)、真实跟踪轨迹数量(mostly tracked targets，MT，用 M_{T} 表示)和丢失轨迹数量(mostly lost targets，ML，用 M_{L} 表示)。目标被误报次数 FP 指跟踪错误的目标数量(跟踪位置和真实位置偏移大于阈值的目标数量)；目标丢失次数 FN 指跟踪失败的目标数量(目标真实位置没有与之对应的跟踪位置)；真实跟踪轨迹数量 MT 指跟踪结果占其真实轨迹长度比例大于 80%的轨迹数量；丢失轨迹数量 ML 指跟踪结果占其真实轨迹长度比例小于 20%的轨迹数量。

MOTA 是一个包含目标被误报次数 FP、目标丢失次数 FN 和身份转换次数 $W_{\mathrm{IDS}}(t)$三个部分的评价多目标跟踪准确度的综合性能指标，其计算式：

$$A_{\mathrm{MOT}}=1-\frac{\sum_t(F_{\mathrm{N}}(t)+F_{\mathrm{P}}(t)+W_{\mathrm{IDS}}(t))}{\sum_t G_{\mathrm{T}}(t)}\ , \qquad (12)$$

其中：t 表示视频帧索引，$F_{\mathrm{N}}(t)$表示第 t 帧中跟踪失败的目标数量，$F_{\mathrm{P}}(t)$表示第 t 帧中跟踪错误的目标数量，$W_{\mathrm{IDS}}(t)$表示第 t 帧中目标轨迹 ID 转换的次数，$G_{\mathrm{T}}(t)$ 表

图 6 多目标跟踪结果展示图。(a) MOT16-01 序列跟踪结果图；(b) MOT16-03 序列跟踪结果图；(c) MOT16-06 序列跟踪结果图

Fig. 6 The results of multi-target tracking chart. (a) MOT16-01 sequence tracking result chart; (b) MOT16-03 sequence tracking result chart; (c) MOT16-06 sequence tracking result chart

示第 t 帧中真实目标的数量。

3.2.2 验证算法

本文算法利用了 R-FCN 检测网络模块和 Siamese 相似度量网络模块，其中 R-FCN 检测网络作为候选模型的重要部分，用于对检测和预测结果进行重新识别分类，Siamese 相似度量网络用于提取健壮的目标外观特征完成轨迹与候选之间数据关联。为了证明算法的有效性，我们在 MOT16 训练集上验证了算法各个模块对多目标跟踪的多个不同的性能指标的影响。基准算法采用卡尔曼滤波器预测目标的新位置，然后通过 IOU 计算预测位置与检测位置的重叠率实现轨迹与候选之间数据关联。在基准算法基础上，依次增加各个网络模块，用 S 表示 Siamese 相似性度量网络模块，R 表示 R-FCN 检测网络模块，实验对比结果见表 1。

表 1 在 MOT16 训练集上验证算法各个模块的有效性

Table 1 Verify the validity of each module of the algorithm on the MOT16 training set

算法	S	R	MOTA/(%)↑	FP↓	FN↓	IDSW↓
基准算法			28.9	2493	75805	686
	√		32.8	4159	69428	452
		√	37.7	10803	57430	537
本文算法	√	√	39.8	6131	59898	**328**

注：最优算法性能指标记为红色，次优算法性能指标记为绿色。

表 1 结果展示，在基准算法中，采用 Siamese 网络模块实现轨迹与候选之间相似度量，可以使得算法对目标具有更好的判别能力，其中 IDSW 指标和 MOTA 指标得到明显改善；通过增加 R-FCN 检测网络模块，并构建基于 R-FCN 检测网络的候选选择模型，目标候选能够较好地从检测和预测结果中被挑选，有效地减少目标漏检问题，与基准算法相比，其中 MOTA 和 FN 指标有较大提升，MOTA 指标提高了 8.8%，FN 指标下降幅度较大。根据表 1 显示，我们在推荐的算法中同时加入 Siamese 网络模块和 R-FCN 检测网络模块，可以获得最优的跟踪表现。

3.2.3 评估算法

与其他五种算法的对比结果见表 2。带星号的方法表示在线跟踪，文献[11,13]属于离线跟踪，五种对比算法的实验数据由相关文献提供。表中加粗数据为最优数据，评价指标箭头向上表示值越大性能越好，箭头向下则相反。对于 MOTA 和 MT，其数值越高越好，而 FP 和 FN 越小越好。

由表 2 可知，本文算法在跟踪准确率 MOTA、真实跟踪轨迹数量 MT、丢失跟踪轨迹数量 ML 和目标漏检 FN 等多个性能指标上处于明显优势，对于复杂场景下检测结果丢失和检测漂移等问题处理能力较强，这主要归功于本文算法建立了候选选择模型，目标候选数据不再仅仅依靠目标检测器提供，可以通过 KF 预测提供的候选框进行数据关联，从而提高了目标数据关联的成功率，减少目标漏检和检测漂移对跟踪结果的影响。此外，为了进一步解决目标丢失、提高跟踪准确率和保证跟踪实时性，每隔 5 帧采用 RANSAC 方法对轨迹进行一次优化处理，利用同一条轨迹上前后帧位置信息，恢复丢失目标位置信息，使跟踪轨迹更加连续更加平滑。

表 3 展示了本文算法与上述其他几种算法的运行时间对比结果，采用跟踪速率(单位为 f/s)(frames per second，FPS)衡量每种跟踪算法的运行速率，即每秒钟可以处理视频序列的帧数，计算方式为处理的视频总帧数除以相应的处理时间。GMMCP、MHT_DAM、CDA_DDAL 以及 AMIR 等几种算法的运行速率相差不大，并且每秒处理速率都不大于一帧，HLSP_T 和本文算法运行速率分别为 4.8 f/s 和 9.7 f/s，在运行速率上更占优势。同时从表 2 中记录的各个算法在多个跟踪性能指标上来看，AMIR、MHT_DAM 和本文算

表 2 MOT16 测试集实验结果对比

Table 2 Comparison of experimental results of MOT16 test set

算法	MOTA/(%)↑	MT/(%)↑	ML/(%)↓	FP↓	FN↓
GMMCP[11]	38.1	8.6	50.9	6607	105315
MHT_DAM[13]	45.8	**16.2**	43.2	6412	91758
HLSP_T[8]*	35.9	8.7	50.1	6412	107918
CDA_DDAL[9]*	43.9	10.7	44.4	6450	95175
AMIR[7]*	47.2	14.0	41.6	**2681**	92856
本文算法*	48.7	15.7	**39.6**	6632	**86504**

注：最优算法性能指标记为红色，次优算法性能指标记为绿色。

表 3 不同算法跟踪速率对比

Table 3 Speed comparison of various tracking algorithms

算法	速度/(f·s^{-1})
GMMCP[11]	0.5
CDA_DDAL[9]*	0.5
MHT_DAM[13]	0.8
AMIR[7]*	1.0
HLSP_T[8]*	4.8
本文算法*	9.7

法的跟踪表现更为突出，在这三种算法之中，本文算法在 MOT16 测试集上的跟踪表现最佳。本文算法之所以能取得如此好的跟踪表现，原因在于加入了候选选择模型，对目标候选框进行了有效筛选，同时采用轻量级网络 SqueezeNet 作为 R-FCN 网络框架的基本模块，大大减少模型的参数，在保证跟踪准确率情况下，平均速度仍然可以达到 9.7 f/s(仅跟踪时间，不包括目标检测时间)。因此本文算法不仅在跟踪表现上有绝对的优势，而且算法也满足基本的实时性要求。

4 结 论

本文提出了一种基于 R-FCN 框架的多候选关联的在线多目标跟踪方法，通过加入候选选择模型，减少对检测结果的依赖，有效地改善了目标漏检和短时间遮挡情况下跟踪问题；采用 Siamese 网络框架实现基于外观特征的相似性度量以提高复杂场景下目标判别能力；用 RANSAC 算法对已有的轨迹进行优化，从而更进一步地解决检测结果可靠性、目标漏检等问题。实验结果表明，与目前已有算法相比，提出的方法在多个跟踪性能指标处于优势。提出的算法不仅可以应用在行人检测与跟踪的场景，还可以推广到包含多个不同类型目标的复杂跟踪场景，应该注意到对于多种不同类型的目标进行数据关联需要采取同一类型目标之间进行数据关联，减少计算复杂度和匹配失败率。然而，在复杂的场景中使用该方法，还存在多方面的挑战，比如目标被大部分遮挡时难以保证目标被稳定地跟踪，这些问题是后续研究的重点。

参考文献

[1] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137–1149.

[2] Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 2016: 779–788.

[3] Liu X, Jin X H. Algorithm for object detection and tracking combined on four inter-frame difference and optical flow methods[J]. *Opto-Electronic Engineering,* 2018, **45**(8): 170665.
刘鑫, 金晅宏. 四帧间差分与光流法结合的目标检测及追踪[J]. 光电工程, 2018, **45**(8): 170665.

[4] Bewley A, Ge Z Y, Ott L, *et al*. Simple online and realtime tracking[C]//*2016 IEEE International Conference on Image Processing* (*ICIP*), Phoenix, 2016: 3464–3468.

[5] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//*2017 IEEE International Conference on Image Processing* (*ICIP*), Beijing, 2017: 3645–3649.

[6] Thoreau M, Kottege N. Improving online multiple object tracking with deep metric learning[Z]. arXiv: 1806.07592v2[cs:CV], 2018.

[7] Sadeghian A, Alahi A, Savarese S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies[Z]. arXiv: 1701.01909[cs:CV], 2017.

[8] Baisa N L. Online multi-target visual tracking using a HISP filter[C]//*13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Funchal, 2018.

[9] Bae S H, Yoon K J. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(3): 595–610.

[10] Milan A, Schindler K, Roth S. Multi-target tracking by discrete-continuous energy minimization[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(10): 2054–2068.

[11] Dehghan A, Assari S M, Shah M. GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking[C]//*2015 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 2015: 4091–4099.

[12] Qi M B, Yue Z L, Shu K, *et al.* Multi-object tracking using hierarchical data association based on generalized correlation clustering graphs[J]. *Acta Automatica Sinica*, 2017, **43**(1): 152–160.
齐美彬, 岳周龙, 疏坤, 等. 基于广义关联聚类图的分层关联多目标跟踪[J]. 自动化学报, 2017, **43**(1): 152–160.

[13] Wen L Y, Li W B, Yan J J, *et al*. Multiple target tracking based on undirected hierarchical relation hypergraph[C]//*2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 2014: 1282–1289.

[14] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks[C]//*2015 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 2015: 4353–4361.

[15] Dai J F, Li Y, He K M, *et al*. R-FCN: Object detection via region-based fully convolutional networks[Z]. arXiv: 1605.06409[cs:CV], 2016.

[16] Iandola F N, Han S, Moskewicz M W, *et al*. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 MB model size[Z]. arXiv: 1602.07360[cs:CV], 2016.

[17] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[Z]. arXiv: 1406.4729[cs:CV], 2014.

[18] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 2016: 770–778.

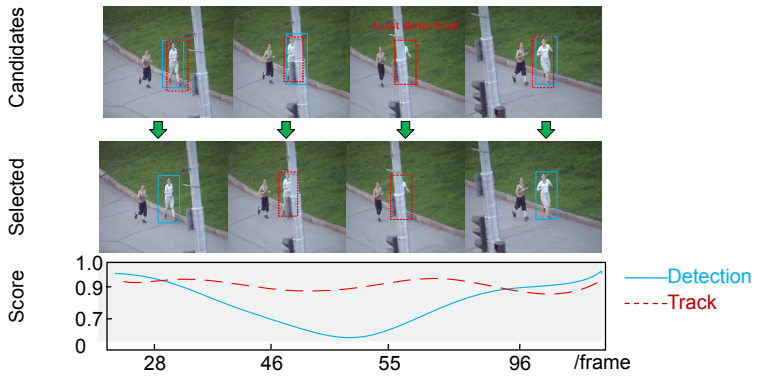
[19] Zheng L, Shen L Y, Tian L, *et al*. Scalable person re-identification: A benchmark[C]//*2015 IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 2015: 1116–1124.

[20] Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the CLEAR MOT metrics[J]. *EURASIP Journal on Image and Video Processing*, 2008, **2008**: 246309.

Multi-candidate association online multi-target tracking based on R-FCN framework

E Gui, Wang Yongxiong*

School of Optical-Electrical and Computer Engineering,
University of Shanghai for Science and Technology, Shanghai 200093, China



Candidates selection flow chart

Overview: As the application basis of human behavior recognition, semantic segmentation and unmanned driving, multi-target tracking is one of the research hotspots in the field of computer vision. In complex tracking scenarios, in order to track multiple targets stably and accurately, many difficulties in tracking need to be considered, such as camera motion, interaction between targets, missed detection and error detection. In recent years, with the rapid development of deep learning, many excellent multi-target tracking algorithms based on detection framework have emerged, which are mainly divided into online multi-target tracking method and offline multi-target tracking method. The multi-target tracking framework process on the basis of detection is as following: the target is detected by the off-line trained target detector, and then the similarity matching method is applied to correlate the detection target. Ultimately, the generated trajectory is continuously used to match the detection result to generate more reliable trajectory. Among them, online multi-target tracking methods mainly include Sort, Deep-sort, SDMT, etc., while offline multi-target tracking methods mainly include network flow model, conditional random field model and generalized association graph model. The offline multi-target tracking methods use multi-frame data information to realize the correlation between the target trajectory and the detection result in the data association process, and can obtain better tracking performance, simultaneously. Unfortunately, those methods are not used to real-time application scenarios. The online tracking methods only use the single-frame data information to complete the data association between the trajectory and the new target which is often unreliable, thus the data association of the lost target will be invalid and the ideal tracking effect cannot be obtained. For purpose of solving the reliability problem of the detection results, an online multi-target tracking method based on R-FCN framework is proposed. Firstly, a candidate model combining Kalman filtering prediction results with detection results is devised. The candidate targets are no longer only from the detection results, which enhances the robustness of the algorithm. Secondly, the Siamese network framework is applied to realize the similarity measurement with respect to the target appearance, and the multiple feature information of the target is merged to complete the data association between multiple targets, which improves the discriminating ability of the target in the complex tracking scene. In addition, on account of the possible missed detection and false detection of the target trajectory in the complex scene, the RANSAC algorithm is used to optimize the existing tracking trajectory so that we can obtain more complete and accurate trajectory information and synchronously the trajectories are more continuous and smoother. Finally, compared to some existing excellent algorithms, the experimental result indicates that the proposed method has brilliant performances in tracking accuracy, the number of lost tracks and target missed detections.

Citation: E G, Wang Y X. Multi-candidate association online multi-target tracking based on R-FCN framework[J]. *Opto-Electronic Engineering*, 2020, **47**(1): 190136

Supported by National Natural Science Foundation of China (61673276, 61703277)

* E-mail: wyxiong@usst.edu.cn