基于深度学习的多目标跟踪关联模型设计

侯建华 [1] 张国帅 [1] 项 俊 [1]

摘 要 近年来, 深度学习在计算机视觉领域的应用取得了突破性进展, 但基于深度学习的视频多目标跟踪 (Multiple object tracking, MOT) 研究却相对甚少, 而鲁棒的关联模型设计是基于检测的多目标跟踪方法的核心. 本文提出一种基于深度神经网络和度量学习的关联模型: 采用行人再识别 (Person re-identification, Re-ID) 领域中广泛使用的度量学习技术和卷积神经网络 (Convolutional neural networks, CNNs) 设计目标外观模型, 即利用三元组损失函数设计一个三通道卷积神经网络, 提取更具判别性的外观特征构建目标外观相似度; 再结合运动模型计算轨迹片间的关联概率. 在关联策略上, 采用匈牙利算法, 首先以逐帧关联方式得到短小可靠的轨迹片集合, 再通过自适应时间滑动窗机制多级关联, 输出各目标最终轨迹. 在 2DMOT2015、MOT16 公开数据集上的实验结果证明了所提方法的有效性, 与当前一些主流算法相比较, 本文方法取得了相当或者领先的跟踪效果.

关键词 多目标跟踪, 深度学习, 度量学习, 关联模型, 多级关联

引用格式 侯建华, 张国帅, 项俊. 基于深度学习的多目标跟踪关联模型设计. 自动化学报, 2020, **46**(12): 2690−2700

DOI 10.16383/j.aas.c180528

Designing Affinity Model for Multiple Object Tracking Based on Deep Learning

HOU Jian-Hua[1] ZHANG Guo-Shuai[1] XIANG Jun[1]

Abstract While deep learning has made a breakthrough in many sub-fields of computer vision recently, there are only a few deep learning approaches to multiple object tracking (MOT). Since the key component in detection based multiple object tracking is to design a robust affinity model, this paper proposes a novel affinity model based on deep neural network and metric learning, that is, metric learning, a widely used technique in the task of person re-identification (Re-ID), is exploited with convolutional neural networks (CNNs) to design the object′s appearance model. Specifically, we adopt a three-channel CNNs that is learned by triplet loss function, to extract the discriminative appearance features and compute appearance similarity between objects. The appearance affinity is then combined with motion model to estimate associating probability among trajectories. A hierarchical association strategy is employed by the Hungarian algorithm. At the low level, a set of short but reliable tracklets are generated in a frame by frame fashion. These tracklets are then further associated to form longer tracklets at the higher levels via an adaptive sliding-window mechanism. Experiment results in the challenging MOT benchmark demonstrate the validity of the proposed method. Compared with several state-of-the-art approaches, our method has achieved competitive or superior performance.

Key words Multiple object tracking (MOT), deep learning, metric learning, affinity model, multi-level association

Citation Hou Jian-Hua, Zhang Guo-Shuai, Xiang Jun. Designing affinity model for multiple object tracking based on deep learning. *Acta Automatica Sinica*, 2020, **46**(12): 2690−2700

多目标跟踪 (Multi object tracking, MOT) 是计算机视觉领域的一个研究热点, 在视频监控、自动驾驶、机器人导航、行为分析等领域发挥着重要作用[1]. 近 10 年来, 随着检测器性能的不断提升[2−4], 基于检测的多目标跟踪算法[1, 5−6] 受到了广泛关注. 这类方法基本流程如下: 由离线检测器提供视频序列每一帧中各目标的位置 (即检测响应), 跟踪算法的任务是将这些检测响应与其对应的目标进行关联, 最终得到每个目标完整的运动轨迹. 基于检测的多目标跟踪包括两个主要模块: 关联模型、关联状态推理 (即优化策略), 本文主要研究前者.

关联模型即目标间亲密度模型, 用以计算下一帧检测响应与当前帧目标之间的连接概率 (或者关联代价), 为关联状态推理提供有效的依据. 特征表达是关联模型设计的核心, 其中最常用的是目标外观特征和目标运动特征. 近 20 年来, 研究者们提出了多种特征用于构建目标的外观模型, 例如颜色直

收稿日期 2018-08-02 录用日期 2019-01-09
Manuscript received August 2, 2018; accepted January 9, 2019
国家自然科学基金 (61671484, 61701548), 湖北省自然科学基金 (2018CFB503), 中南民族大学中央高校基本科研业务费专项资金项目 (CZQ17001, CZZ18001, CZY18046) 资助
Supported by National Natural Science Foundation of China (61671484, 61701548), Hubei Provincial Natural Science Foundation of China (2018CFB503), and Fundamental Research Funds for the Central Universities, South-Central University for Nationalities (CZQ17001, CZZ18001, CZY18046)
本文责任编委 桑农
Recommended by Associate Editor SANG Nong
1. 中南民族大学电子信息工程学院 武汉 430074
1. College of Electronic Information Engineering, South-Central University for Nationalities, Wuhan 430074

方图[7]、方向梯度直方图 (Histogram of oriented gradient, HOG) 特征[8]、协方差特征[9] 等; 外观特征在多目标跟踪过程中发挥着重要的作用, 但是在拥挤场景、以及目标 (例如行人) 具有相似外观的场景, 仅靠外观特征易导致错误关联. 因此很多研究工作通过建模目标动态特性, 将运动特征与外观特征相融合, 构建更鲁棒的目标特征表达[6, 10]. 上述手工设计的特征有力推动了多目标跟踪研究的发展, 但自 2015 年以来更具挑战性的 MOTChallenge[11−12] 数据集的公开, 手工设计的特征已经难以取得令人满意的效果. 例如, 目标间的严重遮挡、剧烈的光照变化和目标形变等, 将可能够导致相同目标之间距离远大于不同目标之间的距离, 造成目标之间的错误关联.

近几年, 深度神经网络 (Deep neural networks, DNNs) 因其强大的特征学习与表达能力, 在图像分类[13]、目标检测[3] 等计算机视觉经典领域的应用取得了突破性进展; 深度学习在视觉目标跟踪 (通常是单目标跟踪) 领域也得到了深入研究[14−16]. 但基于深度学习的多目标跟踪研究却相对甚少, 其主要原因包括[17−18]: 1) 多目标跟踪算法训练样本集少, 难以满足神经网络需要大量训练样本集的要求; 2) 现有的深度神经网络大多是在图像分类数据集上做离线训练基础上得到的, 其缺陷是难以分辨目标间的细微差异、难以捕捉到视频目标中的运动特征. 以下对近年来提出的一些代表性方法做一个简要回顾.

在现有的基于深度学习的多目标跟踪方法中, 一种常见思路是采用孪生卷积神经网络 (Siamese convolutional neural networks, Siamese CNNs) 提取外观特征, 判断两个检测响应是否属于同一条轨迹. Leal-Taixé 等[19] 以一对检测响应的原始图像和光流图像为输入, 由 Siamese CNNs 提取局部时空域特征, 再根据两个检测响应之间的几何及相对位置变化提取上下文信息; 采用梯度增强分类算法, 结合局部特征与上下文信息, 利用对比损失训练整个网络, 得到两个检测响应之间的关联概率; 最后采用匈牙利算法进行数据关联. Sadeghian 等[20] 同时考虑目标外观、运动、以及目标间相互作用机制, 提出了一种基于神经网络和多线索特征融合的关联模型, 利用二元损失函数 (验证损失) 分别训练一个 Siamese CNNs 和一个长短时记忆网络 (Long short-term memory, LSTM) 作为外观模型和运动模型. Tang 等[21] 将行人多目标跟踪任务视为行人再识别问题, 通过融合行人姿态信息, 设计基于 Siamese CNNs 的外观模型. 值得指出的是, 为了改善关联模型的鲁棒性, 近年来在多目标跟踪研究中, 已有将度量学习嵌入到神经网络的报道. Wang 等[22] 提出了一种联合学习卷积神经网络和时域约束度量的轨迹片关联方法, 首先在辅助数据集上离线训练 Siamese CNNs, 用该网络提取两个检测响应的外观特征, 同时对轨迹片施加一个分段时域约束, 构建多任务损失函数, 由一个 Mahalanobis 矩阵和时域约束矩阵组成, 通过同时训练上述两个矩阵, 得到鲁棒的目标外观相似度; 在此基础上, 再结合传统的方法得到运动相似度, 将轨迹片关联问题转换为广义线性分配问题, 采用软分配算法[23] 求其优化解. 此外, Milan 等[24] 将递归神经网络 (Recurrent neural networks, RNNs) 引入到多目标跟踪, 依据贝叶斯滤波理论, 采用 RNNs 对复杂的目标动态特性进行建模, 实现目标状态的预测、更新、以及处理新目标的出现和旧目标消失; 由于在状态更新时需要将下一时刻的观察值与对应的目标相匹配 (即数据关联), 文献 [24] 设计了一个 LSTM 网络来完成此功能; 需要指出的是, 该方法仅依据目标运动特性实现了端到端网络学习和多目标跟踪, 虽然没有达到主流算法的跟踪结果, 但为基于深度学习的视频多目标跟踪的深入研究提供了许多新的、有价值的思路[18, 25−26].

行人再识别 (Person re-identification, Re-ID) 和行人多目标跟踪任务之间存在相似性[27], 即两者都需要判断两个目标是否属于同一人; 但与多目标跟踪不同, Re-ID 问题主要依靠外观特征, 未能有效利用目标的运动特性. 文献 [28] 对 Re-ID 问题中损失函数的研究表明, 三元组损失的性能要优于文献 [20−21] 中使用的二元损失 (验证损失). 受以上启发, 本文将深度卷积神经网络与度量学习相结合, 提出了一种新的多目标跟踪关联模型: 1) 采用 Re-ID 领域中广泛使用的度量学习技术设计外观模型, 即利用三元组损失函数设计一个三通道卷积神经网络, 提取更具判别性的外观特征构建目标外观相似度, 再结合运动模型计算轨迹片间的关联概率; 2) 采用匈牙利算法[29] 通过多级关联生成各目标轨迹. 在 2DMOT2015[12]、MOT16[11] 公开数据集上的实验结果表明, 与当前一些主流算法相比较, 本文方法取得了相当的跟踪效果, 在部分指标上取得了领先.

1 整体算法框架

图 1 给出了本文方法的整体框架. 视频序列输入至目标检测器, 检测器输出每一帧中的目标检测响应.

1) 初级关联: 利用训练好的外观模型提取检测响应的外观特征, 以此计算相邻两帧检测响应之间的外观关联代价矩阵 L_A; 通过计算相邻帧行人检测框中心点坐标的距离获得运动关联代价矩阵 L_M; 由关联代价矩阵 $L = L_A + L_M$ 和匈牙利算法, 采用逐

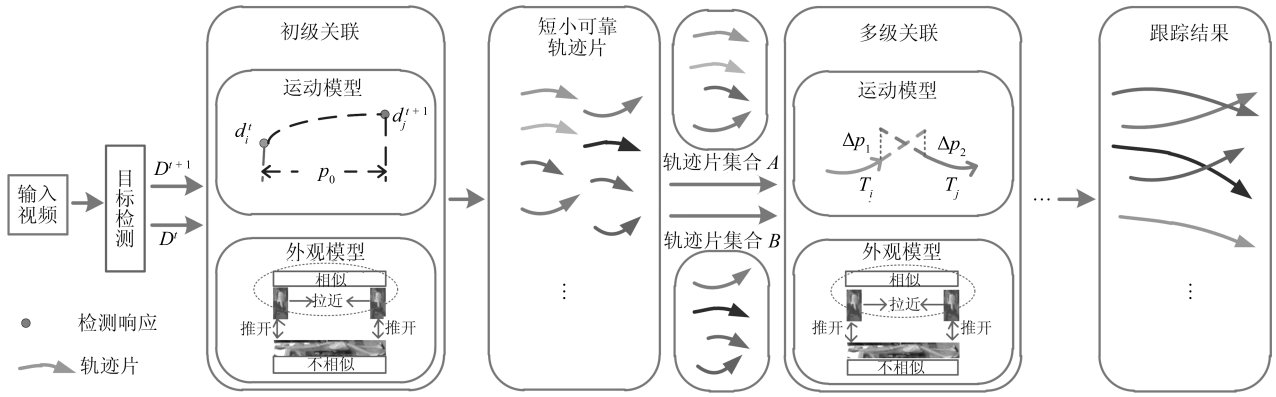


图 1 多目标跟踪方法整体框架

Fig. 1 The overall framework of multi-object tracking method

帧关联的方式得到短小可靠的轨迹片集合.

2) 多级关联: 在初级关联基础上, 依据轨迹片时域约束关系, 采用自适应时间滑动窗机制生成轨迹片集合 A 和 B; 利用关联模型计算两个集合轨迹之间的关联代价, 再由匈牙利算法进行最优关联分配.

2 外观模型

目标外观相似性度量是多目标跟踪关联模型的核心, 也是本文研究的重点. 外观模型的设计要求提取具有判别能力的视觉特征, 且满足正样本 (相同目标) 之间外观关联代价小 (或相似度大)、负样本 (不同目标、或者目标与背景) 之间具有较大的外观关联代价. 上述要求与行人再识别 (Re-ID) 任务是一致的[27], 本文基于深度神经网络、采用在 Re-ID 领域应用广泛的度量学习技术设计外观模型.

2.1 三通道外观模型

本文采用文献 [30] 的方法, 构建三通道卷积神经网络模型, 如图 2 所示, 由三个结构相同、权值参数共享的子网络 Model-A 组成三个通道, 三元组样本分别输入至三个通道, 提取外观特征后, 再由三元损失函数训练整个网络.

子网络 Model-A 结构如下: 在深度残差网络 ResNet-50[31] 的基础上去掉最后一层, 再增加两个全连接层, 其中第一个全连接层节点数为 1 024, 采用批量归一化处理 (Batch normalization)[32] 和 ReLU[33] 激活函数; 第二个全连接层包含 128 个节点, 不使用激活函数, 输出 128 维特征向量. 输入图像尺寸为 256×128 像素.

训练外观模型包括两部分: 首先学习一个映射 $f(x)$, 将输入样本图像 x 映射到嵌入特征空间 (Embedding feature space) $\mathbb{R}^r$[34], r 为嵌入特征空间的维数; 再设计一个度量函数, 使得 $\mathbb{R}^r$ 中正样本对之间的距离远小于负样本对. 在图 2 中, 子网络 Model-A 对应映射 $f(x)$, 即特征提取; 三元损失函数对应度量学习.

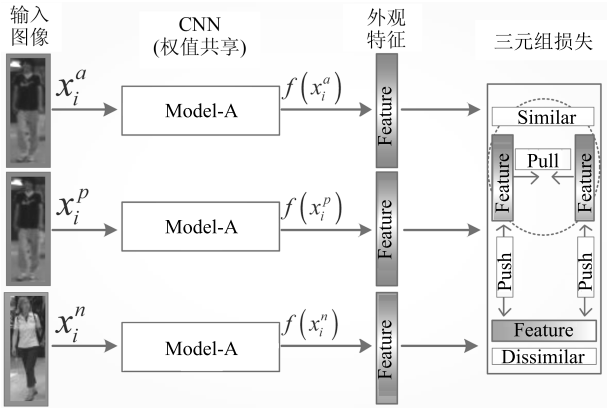


图 2 三通道外观模型训练框图

Fig. 2 Three-channel appearance model training block diagram

设 $Z_i = \langle x_i^a, x_i^p, x_i^n \rangle$ 表示第 i 个三元组样本, Z_i 由三张输入图片构成, 其中, x_i^a 为锚点样本; x_i^p 来自与 x_i^a 相同的目标, x_i^p 与 x_i^a 构成一对正样本; x_i^n 来自与 x_i^a 不同的目标或者背景, x_i^n 与 x_i^a 构成一对负样本. 采用三元组损失训练图 2 所示网络, 其中对三元组样本约束如下[30]:

$$\|f(x_i^a) - f(x_i^p)\|_2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2 \quad (1)$$

设 N 为三元组训练样本总数, 则三元组损失定义为

$$L_{\text{trip}} = \frac{1}{N}\Big(\sum_{i=1}^{N} \|f(x_i^a) - f(x_i^p)\|_2 - \|f(x_i^a) - f(x_i^n)\|_2 + \alpha\Big)_+ \quad (2)$$

其中, $(\cdot)_+ = \max(0, \cdot)$.

三元组损失学习要求所有负样本对的距离要大

于正样本对的距离一个正的最小间隔 α, 即在训练过程中损失不断下降, 使得锚点与正样本越来越接近, 而锚点与负样本距离则越远. α 取值越小, 则在特征空间中锚点与正样本集不需要拉的太近, 锚点与负样本集也不需要拉的太远, 容易满足收敛条件; 但由于正负样本距离没有被拉开, 存在的风险是不能够很好地区分有歧义的数据. 反之, α 取值越大, 可以较有把握地区分较为相似的图像; 但由于在学习过程中要拉近锚点与正样本之间的距离, 同时拉远锚点与负样本之间的距离, 将给网络训练带来一定的困难, 即损失易处于一种很大的状态, 参数更新震荡严重, 模型训练困难. 因此, 设置一个合理的 α 对于基于三元损失的网络训练至关重要. 本文参考文献 [30] 将 α 设置为 1.0.

三元组损失能够在特征空间拉近相同目标之间的距离, 同时增加不同目标之间的距离. 经过三元组损失训练的 CNN, 能够提取具有判别能力的特征, 保证相同目标之间的外观差距远小于不同目标之间的外观差距.

2.2 训练

三元组样本数据的生成是训练外观模型的关键之一. 为保证训练样本的多样性, 本文在多目标跟踪数据库 2DMOT2015[12] 和 MOT16[11] 训练集中的 Ground truth 的基础上, 还使用了行人再识别数据库 Market-1501[35] 和 CUHK03[36]. 需要说明的是, MOT16 训练集中的视频 MOT16-02 作为实验部分的验证集, 不用于本文训练. 由于 MOT 训练数据集 (Ground truth, GT) 和测试数据集中的检测响应 (Detection, DT) 存在一定的差异性, 因此直接用 GT 数据训练网络模型, 在测试集上应用往往得不到理想的精度. 针对此问题, 本文采用在 GT 数据集上人为加入高斯噪声的方法来模拟 DT 的数据产生过程, 即通过适量抖动 GT 行人检测框大小 (即改变中心点位置及尺寸) 的方式, 产生与 DT 尽可能相似的分布数据训练网络模型. 针对存在许多遮挡严重的视频, 本文采用文献 [10] 提出的遮挡推理策略, 通过估计每个运动目标特定时刻的可见度, 将遮挡较大 (可见度小于 0.5) 的目标剔除训练集, 不参与模型训练.

上述训练数据共包括 3 824 个目标, 以 Batch 为单位生成三元组样本. 首先从训练数据中随机选取 5 个不同目标作为锚点 x_i^a, 每一个锚点随机挑选对应的正样本 x_i^p 和负样本 x_i^n, 得到三元组 $\langle x_i^a, x_i^p, x_i^n\rangle$. 每一个目标对应 20 个三元组, 一个 Batch 中包含 100 个三元组, 用于外观网络模型的一次正向传播. 使用 TensorFlow 深度学习框架进行网络训练, 初始化权重来源于文献 [37], 使用 AdamOptimizer[38] 方法更新网络参数, 最小化三元组损失. 初始学习率设置为 0.0001, 每迭代 10 000 次, 学习率衰减 10 %. 实验中迭代 600 000 次时, 网络已趋于收敛.

在测试阶段, 只使用一个子网络 Model-A, 依次输入各帧检测响应, Model-A 输出其外观特征.

3 基于外观和运动特性的轨迹关联

依据外观模型可以得到外观关联代价, 但由于多目标跟踪场景的复杂性, 不同目标也可能具有相似的外观, 此时仅靠外观特征易导致错误关联. 本文在外观模型的基础上, 引入运动模型, 利用目标空域信息提高关联的准确度. 另一方面, 为了降低计算复杂度, 首先进行逐帧初级关联方式, 得到可靠的短小轨迹片; 再采用自适应时间滑动窗对短小轨迹片进行多级关联, 得到完整的目标运动轨迹. 以下分初级关联、以及其后的多级关联两种情况, 分别介绍关联代价的计算.

3.1 初级关联

3.1.1 外观关联代价

依次将各帧检测响应输入至外观网络 Model-A, Model-A 输出对应的外观特征. 设当前帧目标集合为 D^t, 下一帧检测响应集合为 D^{t+1}, 计算两个集合中当前帧 (第 t 帧) 目标 i 的检测响应 d_i^t 和下一帧 (第 $t+1$ 帧) 目标 j 的检测响应 d_j^{t+1} 之间的外观关联代价

$$L_A(d_i^t, d_j^{t+1}) = \left\| f(d_i^t) - f(d_j^{t+1}) \right\|_2 \tag{3}$$

需要说明, 如果当前帧目标为轨迹片, 则 d_i^t 为该轨迹片中置信度最高的检测响应.

3.1.2 运动关联代价

依据检测响应框的中心点坐标, 得到目标和检测响应的位置差, 以此计算运动关联代价

$$L_M(d_i^t, d_j^{t+1}) = \left\| p_i^t - p_j^{t+1} \right\|_2 \tag{4}$$

其中, p_i^t, p_j^{t+1} 分别为检测响应 d_i^t, d_j^{t+1} 的中心点二维坐标矢量. 如果当前帧目标为轨迹片, 则 d_i^t 为该轨迹片中的最后一帧检测响应. 最终的初级关联代价为

$$L(d_i^t, d_j^{t+1}) = L_A(d_i^t, d_j^{t+1}) + L_M(d_i^t, d_j^{t+1}) \tag{5}$$

3.2 多级关联

3.2.1 自适应时间滑动窗机制

通过初级关联得到短小可靠的轨迹片后, 本文使用自适应时间滑动窗机制对其做进一步关联, 最

终输出各目标完整的运动轨迹. 其工作机理是采用自适应滑动时间窗的方式不断构建可关联轨迹片集. 这里可关联轨迹片集是指两个轨迹片集合 A 与 B, 其中, A 定义为当前时刻 t_s 已跟踪目标轨迹集合, B 是滑动窗内待关联的轨迹片集合, 集合 B 的构建满足: 1) B 中的轨迹集与 A 中任意轨迹不存在时间交集 (也即可关联); 2) B 中任意两轨迹片间存在时间交叠.

这里自适应窗口大小的含义是指从 t_s 开始滑动时间窗口, 窗口大小选取为满足集合 B 定义的第 2 个条件的最大值. 设 t_s 时刻 A, B 集如图 3 (a) 所示, 其中集合 A 为已跟踪轨迹, 集合 B 中虚线框的宽表示滑动窗自适应时间跨度, 其选取原则是恰好使得 B 集合里的所有轨迹片集满足 "任意两两轨迹片彼此存在时间交叠", 如果继续增加滑动窗大小, 也即 B 虚线宽增加, 将 11 帧和 12 帧轨迹片包含到集合 B, 则 B 中轨迹片 T_1 与第 11 帧轨迹片 T_2 不满足集合 B 的第 2 个约束条件.

在当前滑动窗机制下生成 A, B 集以后, 采用匈牙利算法实现可关联轨迹片之间的最优关联状态. 如图 3 (b) 所示, 将 B 中被关联的轨迹片与 A 中轨迹串联并加入到集合 A (例如图中虚线所示), 同时 B 中没有关联上的轨迹片也作为新目标轨迹加入 A 集 (例如图中轨迹片 T_3), 至此完成了集合 A 的更新. 下一时刻的滑动窗从 $t_s = 10$ 时刻开始, 重新按照自适应滑动方式构建新的集合 B, 直至滑动完整个视频集, 算法结束.

采用自适应滑动窗技术可以部分解决遮挡问题, 假设因为遮挡导致某目标检测响应丢失, 但自适应滑动窗可以考虑具有时间跨度的轨迹片间的相关性, 即利用该目标在无遮挡阶段的检测数据进行关联. 但基于滑动窗的数据处理模式存在一定的时间延迟, 不能实现逐帧实时输出, 无法满足对实时性要求较高的应用场景.

3.2.2 外观关联代价

设 T_i, T_j 分别代表集合 A, B 的一条轨迹片, 其外观关联代价为

$$L_A(T_i, T_j) = \|f(T_i) - f(T_j)\|_2 \tag{6}$$

其中, $f(T_i)$, $f(T_j)$ 对应轨迹片中置信度最高的检测响应的外观特征.

3.2.3 运动关联代价

采用文献 [39] 的方法构建多级关联的运动模型. 首先定义轨迹片运动关联概率为

$$g(T_i, T_j) = G(\Delta p_1, \Sigma)G(\Delta p_2, \Sigma) \tag{7}$$

其中, $G(\cdot, \Sigma)$ 代表零均值高斯分布, Σ 为高斯分布的方差.

多级关联运动模型如图 4 所示, 设 p_j^{head} 为轨迹片 T_j 第一帧检测框的中心点位置 (对应时刻记为 t_j^{head}), v_j^{head} 表示 T_j 所对应目标 j 的运动速度; p_i^{tail} 代表轨迹片 T_i 最后一帧检测框的中心点位置 (对应时刻记为 t_i^{tail}), v_i^{tail} 是轨迹片 T_i 所对应目标 i 的运动速度. 假设目标做匀速运动, $\Delta t = t_j^{\text{head}} - t_i^{\text{tail}}$ 表示 T_j 第一帧响应与 T_i 最后一帧响应的时间间隔, 则目标 i 在时刻 t_j^{head} 的预测位置为 $p_i^{\text{tail}} + v_i^{\text{tail}} \cdot \Delta t$, 该预测值与待关联目标 j 实际位置的误差为

$$\Delta p_1 = p_i^{\text{tail}} + v_i^{\text{tail}} \times \Delta t - p_j^{\text{head}} \tag{8}$$

同理,

$$\Delta p_2 = p_j^{\text{head}} - v_j^{\text{head}} \times \Delta t - p_i^{\text{tail}} \tag{9}$$

假定 Δp_1 和 Δp_2 服从高斯分布, 则预测位置与待关联目标实际位置的误差越小, 对应的轨迹片之间的运动相似度就越大.

由式 (7) 计算运动关联概率, 再由下式得到对应的运动关联代价为

$$L_M(T_i, T_j) = -\ln(g(T_i, T_j)) \tag{10}$$

最终的关联代价为

$$L(T_i, T_j) = L_A(T_i, T_j) + L_M(T_i, T_j) \tag{11}$$

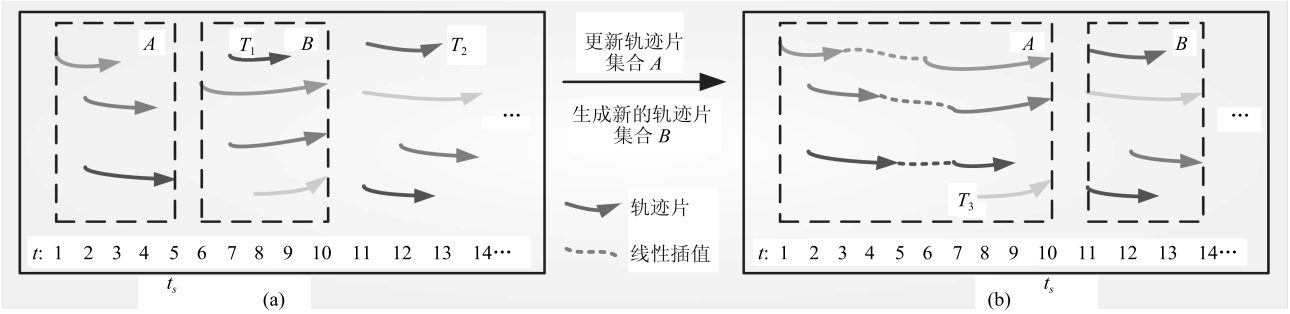


图 3 自适应时间滑动窗原理示意图

Fig. 3 Diagram of adaptive time sliding window principle

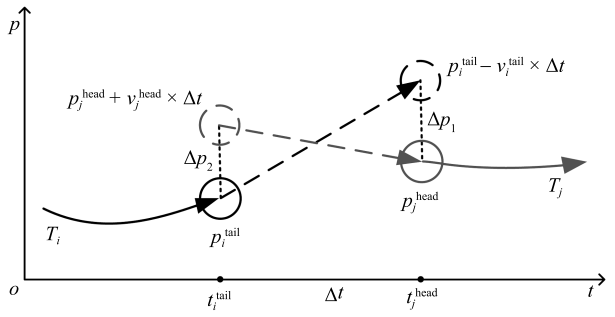


图 4　多级关联中的运动模型示意图

Fig. 4　Diagram of motion model in multi-level association

4　实验结果及分析

本节首先介绍多目标跟踪性能评价指标; 然后通过剥离对比实验 (Ablation study) 证明外观模型、运动模型及三元组损失的有效性; 同时, 在行人数据库 MOTChallenge[11−12] 测试集上与当前主流跟踪算法进行了比较分析, 并给出了直观的可视化跟踪结果. 最后在 UA-DETRAC 车辆数据集上, 进一步验证本文算法的有效性. 本文实验的硬件环境: CPU 为 Intel Xeon(R) E5-2620@2.10 GHz, 内存为 128 GB.

4.1　评价指标

按照国际通行方法, 采用文献 [11, 40] 定义的标准来评估多目标跟踪算法的性能, 包括:

MT (↑): 大部分被跟踪的目标. 针对某一个目标, 实际轨迹长度占该目标真实轨迹长度的百分比大于 80 %, 则认为该目标被大部分跟踪.

ML (↓): 大部分未被跟踪的目标. 针对某一个目标, 实际轨迹长度占该目标真实轨迹长度的百分比小于 20 %, 则认为该目标大部分未被跟踪.

FP (↓): 被预测为正的负样本, 即虚检.

FN (↓): 被预测为负的正样本, 即漏检.

Frag (↓): 轨迹断裂的次数. 即还存在能关联, 而没有被关联上的短小轨迹片.

IDS (↓): 针对某一条轨迹, 其对应的目标身份发生变化的次数.

MOTA (↑): 多目标跟踪准确率. 综合 FP、IDS 和 FN 计算, 是衡量多目标跟踪算法性能的主要指标.

MOTP (↑): 多目标跟踪精度. 根据检测响应与真实数据的行人框的重合率计算, 是衡量多目标跟踪算法性能的主要指标.

在上述指标中, ↑ 表示该指标越大性能越好, ↓ 则表示该指标越小性能越好.

4.2　剥离对比实验及分析

为了验证本文外观模型、运动模型在多目标跟踪中的作用, 以 MOT16-02 作为验证集, 设计了剥离实验 (Ablation study). 同时为了比较三元组损失与常规的二元损失的差异性, 我们在不改变外观模型网络结构的前提下, 设计并使用 Siamese CNN, 计算一对检测响应之间的关联概率, 定义二元交叉熵损失来训练外观模型.

实验结果如表 1 所示, 其中 A 表示外观模型, M 表示运动模型, T 表示使用三元组损失, V 表示使用二元交叉熵损失, 加粗数据表示最好结果. 从表 1 可以看出:

1) 外观特征在多目标跟踪过程中发挥主要作用. 在使用三元组损失时, 仅使用运动特征时, MOTA 为 17.6 %; 而仅使用外观特征时, MOTA 为 19.5 %. 上述结果与实际情况是相符的, 因为在现实场景中, 人眼主要通过目标衣着的颜色、形状等外观视觉特征来进行跟踪; 同时卷积神经网络 (CNN) 具有很好的提取视觉结构化信息的能力.

2) 外观和运动特征相结合能够有效提升跟踪精度. 同时使用外观和运动特征, 得到的 MOTA 最高, 为 21 %. 在复杂场景下, 目标之间易出现遮挡; 另一方面, 距离较远的不同目标之间也有可能具有相同的外观. 针对这些情况, 仅使用外观特征不能得到准确的跟踪. 本文结合运动模型, 利用位置信息对关联进行约束, 与外观特征协同发挥作用.

3) 与二元损失 (验证损失) 相比, 使用度量学习技术中的三元组损失能够显著提升跟踪精度,

表 1　剥离对比实验结果

Table 1　Results of ablation study

Trackers	MOTA (↑)	MOTP (↑)	MT (↑) (%)	ML (↓) (%)	FP (↓)	FN (↓)	IDS (↓)
A + T	19.5	74.6	7.41	66.70	109	14 202	43
M + T	17.6	74.6	7.40	**64.80**	307	14 326	70
A + M + T	**21.0**	74.3	**9.26**	70.40	175	**13 893**	**16**
A + M + V	14.7	**75.1**	1.85	67.00	**60**	14 804	339

MTOA 由 14.7 % 上升到 21.0 %, 说明了将行人再识别领域中的三元组损失应用于多目标跟踪的可行性.

4.3 与主流算法的对比

表 2 给出了本文算法与当前主流多目标跟踪算法在 MOT16 测试集上的结果对比. 从表 2 可以看出, 本文算法在 MOTA、MT、FP、FN 指标上优于之前文献 [41−44] 中的跟踪算法. 本文算法得到的 MOTA 为 43.1 %, 低于文献 [20] 的 AMIR 方法. 文献 [20] 同时考虑了目标外观、运动、以及目标间相互作用机制, 并且采用 LSTM 网络建模目标在时域的非线性运动特性, 而本文方法仅采用了传统的线性运动模型, 且未能提取时域运动信息, 这也是今后待改进之处. 另外一点值得说明的是, 本文算法的 IDS 优于文献 [17, 20, 41−42, 44−45] 的方法, IDS 越小, 说明跟踪轨迹的目标身份变化次数少, 跟踪结果的可靠性高.

表 3 列出了在 2DMOT2015 测试集上的结果对比. 从表 3 可以看出, 与基于孪生卷积神经网络 (Siamese CNNs) 计算关联概率的算法[19, 22] 相比, 本文算法在 MOTA 上具有较大的优势. 本文方法框架与文献 [22] 类似, 都采用度量学习技术和神经网络提取外观特征, 同时利用了传统的运动模型, 最后用匈牙利算法进行轨迹关联; 但不同的是本文采用三元组损失训练网络, 用一个 CNN 提取外观特征. 而文献 [22] 采用分段时域约束构建多任务损失函数训练 Siamese CNNs, 并用于外观特征提取. 同时, 本文三元组损失只对类间差距有约束, 而对类内差距未做要求, 因此文献 [22] 训练过程中的约束条件更强. 但本文方法在实际中能获得更好的跟踪性能, 与文献 [22] 相比, 本文得到的 MOTA 提高了 4.6 个百分点 (从 29.6 % 到 34.2 %), 说明了在这种遮挡严重的场景下, 适当放松训练中的约束条件, 可以使得模型更容易收敛, 在测试时能够获得理想的跟踪效果. 同时也说明了基于三元损失函数的三通道卷积

表 2 MOT16 测试集结果

Table 2 Results of MOT16 test set

Trackers	Mode	MOTA (↑)	MOTP (↑)	MT (↑) (%)	ML (↓) (%)	FP (↓)	FN (↓)	IDS (↓)	HZ (↑)
AMIR[20]	Online	**47.2**	75.8	14.0	**41.6**	**2 681**	**92 856**	774	1.0
CDA[45]	Online	43.9	74.7	10.7	44.4	6 450	95 175	676	0.5
本文	Online	43.1	74.2	12.4	47.7	4 228	99 057	495	0.7
EAMTT[41]	Online	38.8	75.1	7.9	49.1	8 114	102 452	965	**11.8**
OVBT[42]	Online	38.4	75.4	7.5	47.3	11 517	99 463	1 321	0.3
Quad-CNN[17]	Batch	44.1	**76.4**	**14.6**	44.9	6 388	94 775	745	1.8
LIN1[43]	Batch	41.0	74.8	11.6	51.3	7 896	99 224	**430**	4.2
CEM[44]	Batch	33.2	75.8	7.8	54.4	6 837	114 322	642	0.3

表 3 2DMOT2015 测试集结果

Table 3 Results of 2DMOT2015 test set

Trackers	Mode	MOTA (↑)	MOTP (↑)	MT (↑) (%)	ML (↓) (%)	FP (↓)	FN (↓)	IDS (↓)	HZ (↑)
AMIR[20]	Online	**37.6**	71.7	15.8	**26.8**	7 933	**29 397**	1 026	1.9
本文	Online	34.2	71.9	8.9	40.6	7965	31665	794	0.7
CDA[45]	Online	32.8	70.7	9.7	42.2	**4 983**	35 690	614	2.3
RNN_LSTM[24]	Online	19.0	71.0	5.5	45.6	11 578	36 706	1 490	**165.2**
Quad-CNN[17]	Batch	33.8	**73.4**	12.9	36.9	7 879	32 061	703	3.7
MHT_DAM[46]	Batch	32.4	71.8	**16.0**	43.8	9 064	32 060	435	0.7
CNNTCM[22]	Batch	29.6	71.8	11.2	44.0	7 786	34 733	712	1.7
Siamese CNN[19]	Batch	29.0	71.2	8.5	48.4	5 160	37 798	639	52.8
LIN1[43]	Batch	24.5	71.3	5.5	64.6	5 864	40 207	**298**	7.5

神经网络提取外观特征的有效性.

4.4　可视化跟踪结果

为了从直观上说明算法性能, 图 5～8 给出了本文算法在部分视频上的可视化跟踪结果.

图 5 所示是 MOT16-01 的一段视频, 该视频拍摄光线较暗, 背景嘈杂; 其中的目标 7 在从右到左行走时虽然发生了多次遮挡, 但始终都能够被准确跟踪.

图 6 是 MOT16-03 的一段视频, 该视频行人较多, 导致行人之间易发生频繁遮挡, 且不同目标之间会有相同的衣着. 通过观察可以发现, 在这种困难场景下, 虽然产生许多 Frag, 但是我们的算法能够进行保守关联, 将关联错误减少到最小.

图 7 是 MOT16-07 的一段在摄像头抖动严重的条件下拍摄的视频, 本文方法在摄像头严重抖动的条件下, 也能够得到正确的目标轨迹.

图 8 是 MOT16-06 的一段光照条件不断发生

图 5　MOT16-01 跟踪结果 (从左到右依次为第 121、174、248 帧)

Fig. 5　Tracking results of MOT16-01 (121st, 174th, 248th frames from left to right)

图 6　MOT16-03 跟踪结果 (从左到右依次为第 249、307、424 帧)

Fig. 6　Tracking results of MOT16-03 (249th, 307th, 424th frames from left to right)

图 7　MOT16-07 跟踪结果 (从左到右依次为第 397、455、500 帧)

Fig. 7　Tracking results of MOT16-07 (397th, 455th, 500th frames from left to right)

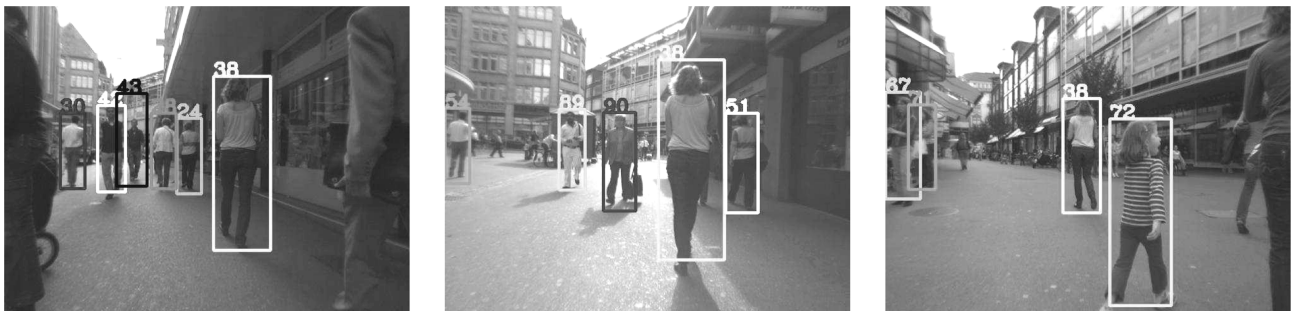


图 8　MOT16-06 跟踪结果 (从左到右依次为第 537、806、1188 帧)

Fig. 8　Tracking results of MOT16-06 (537th, 806th, 1188th frames from left to right)

表 4 UA-DETRAC 数据集跟踪结果

Table 4 Tracking results of UA-DETRAC dataset

	MOTA (↑)	MOTP (↑)	MT (↑) (%)	ML (↓) (%)	FP (↓)	FN (↓)	IDS (↓)
车辆跟踪	65.3	78.5	75.0	8.3	1 069	481	27

图 9 MVI_20032 跟踪结果 (从左到右依次为第 332、360、423 帧)

Fig. 9 Tracking results of MVI_20032 (332nd, 360th, 423rd frames from left to right)

图 10 MVI_39771 跟踪结果 (从左到右依次为第 1、54、113 帧)

Fig. 10 Tracking results of MVI_39771 (1st, 54th, 113th frames from left to right)

变化的视频, 且目标尺寸也在变化, 增加了跟踪难度, 但目标 38 从第 537 帧到第 1 188 帧一直被正确跟踪.

4.5 UA-DETRAC 数据集跟踪实验

为进一步说明本文算法的有效性, 特补充了一组在 UA-DETRAC 数据集上的车辆跟踪实验. 实验中使用 UA-DETRAC 数据集 Train 中的 MVI_20032 和 MVI_39771 作为验证集, R-CNN (Region-CNN)[47] 提供检测响应. 测试时, 直接使用本文的关联模型实现数据关联, 并未进行任何网络参数微调. 采用文献 [11, 40] 定义的标准来评估跟踪算法的性能. 以上两个视频分别在晴天早上和阴天傍晚进行拍摄, 长度分别为 437 帧和 570 帧. 实验结果如表 4 所示, 跟踪可视化结果如图 9 和图 10 所示. 通过定量和定性的结果可以看出, 本文算法在跟踪汽车实验中具有一定的有效性.

5 结论

在基于检测的多目标跟踪框架下, 提出了一种基于深度神经网络和度量学习的关联模型: 利用三元组损失函数设计一个三通道卷积神经网络, 提取更具判别性的外观特征构建目标外观相似度; 再结合运动模型计算轨迹片间的关联代价. 在此基础上, 采用多级关联策略和匈牙利算法, 得到运动目标的完整轨迹. 对各种场景下跟踪结果的定量与定性分析证明了基于三元组损失函数的三通道外观模型的有效性; 在 2DMOT2015、MOT16 公开数据集上与当前一些主流算法进行了比较, 本文方法取得了相当或者领先的跟踪效果. 同时, 通过在 UA-DETRAC 数据集上的跟踪实验说明本文算法在除行人外的其他类别物体 (如汽车) 上也具有一定的有效性. 但本文方法在出现长时间目标遮挡时产生错误关联, 今后可以考虑构建基于递归神经网络的运动特征提取模型, 将其提取的运动特征与原来的外观特征一起送入基于三元组损失的度量网络进行学习, 得到融合特征; 在测试阶段利用融合特征计算轨迹片关联代价, 实现轨迹片关联.

References

1 Luo W H, Xing J L, Milan A, Zhang X Q, Liu W, Zhao X W, et al. Multiple object tracking: A literature review. arXiv preprint arXiv: 1409.7618, 2014.

2 Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(9): 1627−1645

3 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago,

Chile: IEEE, 2015. 1440−1448

4 Yin Hong-Peng, Chen Bo, Chai Yi, Liu Zhao-Dong. Vision-based object detection and tracking: A review. *Acta Automatica Sinica*, 2016, **42**(10): 1466−1489
(尹宏鹏, 陈波, 柴毅, 刘兆栋. 基于视觉的目标检测与跟踪综述. 自动化学报, 2016, **42**(10): 1466−1489)

5 Xiang J, Sang N, Hou J H, Huang R, Gao C X. Hough forest-based association framework with occlusion handling for multi-target tracking. *IEEE Signal Processing Letters*, 2016, **23**(2): 257−261

6 Yang B, Nevatia R. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 1918−1925

7 Nummiaro K, Koller-Meier E, Van Gool L. An adaptive color-based particle filter. *Image and Vision Computing*, 2003, **21**(1): 99−110

8 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 886−893

9 Tuzel O, Porikli F, Meer P. Region covariance: A fast descriptor for detection and classification. In: Proceedings of the 2006 European Conference on Computer Vision. Graz, Austria: Springer, 2006. 589−600

10 Xiang J, Sang N, Hou J H, Huang R, Gao C X. Multitarget tracking using Hough forest random field. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, **26**(11): 2028−2042

11 Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K. MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv: 1603.00831, 2016.

12 Leal-Taixé L, Milan A, Reid I, Roth S, Schindler K. MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv: 1504.01942, 2015.

13 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: MIT, 2012. 1097−1105

14 Guan Hao, Xue Xiang-Yang, An Zhi-Yong. Advances on application of deep learning for video object tracking. *Acta Automatica Sinica*, 2016, **42**(6): 834−847
(管皓, 薛向阳, 安志勇. 深度学习在视频目标跟踪中的应用进展与展望. 自动化学报, 2016, **42**(6): 834−847)

15 Bertinetto L, Valmadre J, Henriques J F, Vedaldi A, Torr P H S. Fully-convolutional siamese networks for object tracking. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 850−865

16 Danelljan M, Robinson A, Khan F S, Felsberg M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 472−488

17 Son J, Baek M, Cho M, Han B. Multi-object tracking with quadruplet convolutional neural networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 3786−3795

18 Emami P, Pardalos P M, Elefteriadou L, Ranka S. Machine learning methods for solving assignment problems in multi-target tracking. arXiv preprint arXiv: 1802.06897, 2018.

19 Leal-Taixé L, Canton-Ferrer C, Schindler K. Learning by tracking: Siamese CNN for robust target association. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA: IEEE, 2016. 418−425

20 Sadeghian A, Alahi A, Savarese S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 300−311

21 Tang S Y, Andriluka M, Andres B, Schiele B. Multiple people tracking by lifted multicut and person re-identification. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 3701−3710

22 Wang B, Wang L, Shuai B, Zuo Z, Liu T, Chan K L, Wang G. Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA: IEEE, 2016. 368−393

23 Gold S, Rangarajan A. Softmax to softassign: Neural network algorithms for combinatorial optimization. *Journal of Artificial Neural Networks*, 1996, **2**(4): 381−399

24 Milan A, Rezatofighi S H, Dick A, Schindler K, Reid I. Online multi-target tracking using recurrent neural networks. In: Proceedings of the 2017 AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2017. 2−4

25 Beyer L, Breuers S, Kurin V, Leibe B. Towards a principled integration of multi-camera re-identification and tracking through optimal Bayes filters. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE, 2017. 1444−1453

26 Farazi H, Behnke S. Online visual robot tracking and identification using deep LSTM networks. In: Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, Canada: IEEE, 2017. 6118−6125

27 Kuo C H, Nevatia R. How does person identity recognition help multi-person tracking? In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2011. 1217−1224

28 Xiao Q Q, Luo H, Zhang C. Margin sample mining loss: A deep learning based method for person re-identification. arXiv preprint arXiv: 1710.00478, 2017.

29 Huang C, Wu B, Nevatia R. Robust object tracking by hierarchical association of detection responses. In: Proceedings of the 2008 European Conference on Computer Vision. Marseille, France: Springer, 2008. 788−801

30 Cheng D, Gong Y H, Zhou S P, Wang J J, Zheng N N. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 1335−1344

31 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770−778

32 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: ACM, 2015. 448−456

33 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: JMLR, 2011. 315−323

34 Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 815−823

35 Zheng L, Shen L Y, Tian L, Wang S J, Wang J D, Tian Q. Scalable person re-identification: A benchmark. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1116−1124

36 Li W, Zhao R, Xiao T, Wang X G. DeepReID: Deep filter pairing neural network for person re-identification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 152−159

37 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv preprint arXiv: 1703.07737, 2017.

38 Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980, 2014.

39 Yang B, Nevatia R. Multi-target tracking by online learning a CRF model of appearance and motion patterns. *International Journal of Computer Vision*, 2014, **107**(2): 203−217

40 Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008, **2008**: 246309

41 Sanchez-Matilla R, Poiesi F, Cavallaro A. Online multi-target tracking with strong and weak detections. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 84−99

42 Ban Y T, Ba S, Alameda-Pineda X, Horaud R. Tracking multiple persons based on a variational Bayesian model. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 52−67

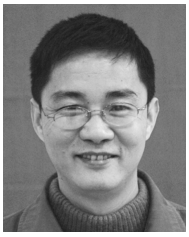
43 Fagot-Bouquet L, Audigier R, Dhome Y, Lerasle F. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 774−790

44 Milan A, Roth S, Schindler K. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(1): 58−72

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

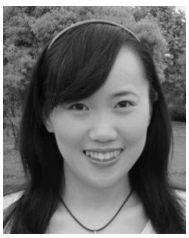
46 Kim C, Li F X, Ciptadi A, Rehg J M. Multiple hypothesis tracking revisited. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 4696−4704

47 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 580−587



侯建华 中南民族大学电子信息工程学院教授. 2007 年获华中科技大学模式识别与智能系统博士学位. 主要研究方向为计算机视觉与模式识别.
E-mail: zil@scuec.edu.cn
(**HOU Jian-Hua** Professor at the College of Electronic Information Engineering, South-Central University for Nationalities. He received his Ph. D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology in 2007. His research interest covers computer vision and pattern recognition.)

张国帅 中南民族大学电子信息工程学院硕士研究生. 2016 年获长春大学学士学位. 主要研究方向为图像处理与模式识别. E-mail: guoshuaiz@scuec.edu.cn
(**ZHANG Guo-Shuai** Master student at the College of Electronic Information Engineering, South-Central University for Nationalities. He received his bachelor degree from Changchun University in 2016. His research interest covers image processing and pattern recognition.)



项 俊 中南民族大学电子信息工程学院讲师. 2016 年获华中科技大学控制科学与工程博士学位. 主要研究方向为计算机视觉与模式识别. 本文通信作者.
E-mail: junxiang@scuec.edu.cn
(**XIANG Jun** Lecturer at the College of Electronic Information Engineering, South-Central University for Nationalities. She received her Ph. D. degree in control science and engineering from Huazhong University of Science and Technology in 2016. Her research interest covers computer vision and pattern recognition. Corresponding author of this paper.)