多目标跨摄像头跟踪技术

焦珊珊，李云波，陈佳林，潘志松

（陆军工程大学，江苏　南京　210000）

[摘　要]　多目标跨摄像头跟踪，指的是对于多个被寻目标，在不同场景的摄像头中找到他们出现的时间和空间信息并加以关联。算法综合利用人脸检测、行人检测、目标跟踪、人脸识别和行人再识别等技术，形成一个统一的技术框架。无论是在实现框架上，还是在准确率上都能达到商用水平。在南京市举办的全球人工智能应用大赛中获得了二等奖。

[关键词]　多目标跨摄像头跟踪；人脸检测；行人检测；目标跟踪；人脸识别；行人再识别

[中图分类号]　E933　　[文献标识码]　A　　[文章编号]　1671－4547（2019）06－0033－09

DOI：10.13943/j.issn 1671－4547.2019.06.08

引言

多目标跨摄像头跟踪项目主要解决跨摄像头跨场景下行人的识别与检索。该技术可以作为人脸识别技术的重要补充，对无法获取清晰拍摄人脸的行人进行跨摄像头连续跟踪，增强数据的时空连续性[1]。原来由人坐在监控室或者翻看监控视频进行检查的工作模式费人费力，该技术可以广泛应用于视频监控、智能安保、刑事侦查等领域，通过提供感兴趣目标的人脸图片或行人姿态，搜索该目标出现的时间和地点，并对检测到的目标在视频的帧序列中进行跟踪，分析跟踪的目标轨迹[2－3]。同时在智能商业方面，通过研究客户运动轨迹和停留时间，获取客户需求，实现客户的私人订制。人工智能即将从“刷脸”跨越到“识人”的新时代。

本算法的研究基于多摄像头跨区域目标的识别和跟踪，输入为给定 4 个室内摄像头拍摄的各 5 个视频，每段视频 3 分钟；4 段室外视频，每段视频两小时；或 200 张感兴趣的人脸图像。要求判别 200 张感兴趣的人脸是否在这些视频中出现，以及出现在视频中的所有行人时间和位置。问题的核心是跨摄像头跟踪，即对在不同摄像头中出现的同一目标进行关联，主要涉及两个方面的识别：人脸识别和行人再识别。该问题的解决还需依赖多种信息和技术，如目标检测、目标跟踪和时空关系等。根据用户需求在室内视频中进行人脸识别，并正确锁定对应行人，再根据行人特征，关联所有室内外视频中该行人的出现时间（帧号）和位置区域。

一、算法实现思路

针对待解决的任务，我们构建了整体的实现框架，主要思路如下图所示：

首先，对室内视频进行人脸检测和人脸识别；然后，根据识别出的感兴趣人脸，获得该行人在视频中出现的全身位置信息（该步骤涉及行人检测）；最后，在室内外视频中再识别该行人，获得时间和位置信息。该过程中，同一人在时间上的关联由目标跟踪完成。

总体来说，实现框架可分解为以下关键技术：人脸检测、行人检测、目标跟踪、人脸识别、人脸行人关联和行人再识别。

[收稿日期]　2019－06－18

[作者简介]　焦珊珊，女，博士研究生，研究方向：计算机视觉、人工智能；
潘志松，男，教授，博士，研究方向：机器学习、人工智能。

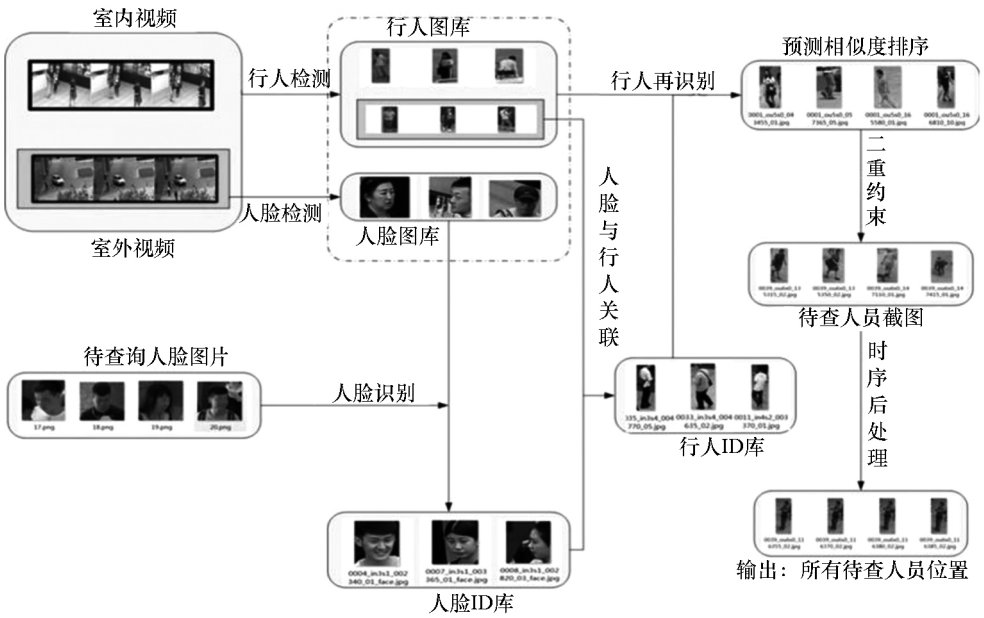

图 1　多目标跨摄像头跟踪流程图

二、方法步骤

（一）人脸检测

采用目前人脸检测效果最优的 PyramidBox[4] 方法，该方法曾在世界上最权威的人脸检测公开评测集 WIDER Face 上达到最高的人脸检测精度。

我们在任务中发现，感兴趣的人脸图片中存在大量分辨率过低，运动模糊以及大部分遮挡的样本。图 2 中我们给出了本次任务中，传统人脸检测方法无法检测定位的部分样例，传统算法对样本描述能力不足，导致后续识别精度受到严重影响。

图 2　实际任务中人脸的模糊、低分辨率与遮挡现象

分析原因如下：一、目前主流的人脸检测数据集都普遍采用高分辨率的图像，其数据集中困难样本不充足；二、目前人脸检测数据集都是基于图像进行检测的，而视频本质上存在前后帧时序关系，丢掉视频的时序关系，用图像的方法检测人脸丢失了很多信息。

针对以上两个问题，解决方法如下：

在目前人脸检测数据集上通过降低分辨率和模糊图像的方式加入困难样本，然后重新训练深度网络模型，进一步提高鲁棒性，网络基本结构采用的方法如图 4，该方法是目前 FDDB and WIDER FACE 数据集上最先进的人脸检测方法，可进一步提高人脸检测的精度。

针对视频中人脸检测的问题，我们在检测的同时对人脸同步进行跟踪，为不属于同一个目标的人脸的后续判别提供重要依据，该方法将在跟踪部分一起介绍。

（二）行人检测

采用目前最佳的目标检测算法 YOLO V3[5]，方法采用多任务多损失训练，在 COCO 数据集上实现了最佳精度，并在视频检测上达到实时效果。针对目标任务室内外清晰度、亮度、透视角度存在差异，而行人形态、树干遮挡以及相近标志都影响了行人检测的效果的问题，我们在旷视发布的 CrowdHuman 行人检测数据集中进行 fine－tune 训练，大幅提高了行人的辨识准确度。

此外，室内视频中存在大量显示屏，显示屏中的行人是蓝色图片，如图 3 所示。背景和行人几乎全是蓝色的，显然不同于真实的行人。我们使用 python 的 opencv 读入图片后比较 RGB 三通道的均值，发现虚假图片的 B 通道均值明显低于

图3　部分蓝光图像

80，而正常行人的三通道基本分布均匀，因此，在行人检测的同时去除了 B 通道均值小于 80 的图片。

（三）目标跟踪

我们提出采用中心位置变化率和特征相似度完成目标前后帧关联匹配的方法，实现了基于检测的目标跟踪。采用基于检测的跟踪思路，首先检测出人脸或行人在前后帧中的位置区域，通过比对前后帧的目标中心位置 p_i 变化率和目标外观特征 f_i 相似度来进行关联匹配，实现跟踪，如图 4 所示。若前后帧目标中心位置变化率和特征相似度均超过设定阈值，则判定目标为同一目标。

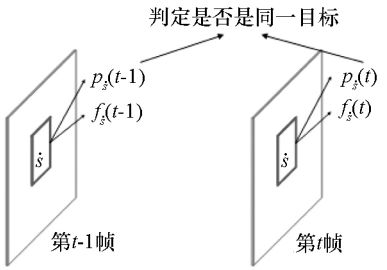

图4　目标跟踪基本思路。其中 p_i 表示目标中心位置坐标，f_i 表示目标外观的特征值，t 表示时刻。

任务主要涉及两类目标：人脸和行人。针对人脸，由前后帧人脸检测区域中心点距离和人脸区域均值的比值作为中心位置变化率；针对行人，由前后帧行人检测区域中心点距离和行人宽度均值的比值作为中心位置变化率。同时提取人脸 68 点特征表示和行人重识别特征表示，特征度量时先对特征进行 L2 归一化，再计算特征余弦距离作为外观特征相似度。

针对给定的视频，同时进行人脸跟踪和行人跟踪。具体实施时，针对前后帧检测得到的人脸和行人区域以及提取的人脸和行人特征进行中心位置变化率计算和特征相似度计算。

（1）人脸跟踪。若中心位置变化率小于阈值 1，且特征欧式距离（特征归一化后等价于余弦相似度）小于 0.6，则判定前后帧目标为同一目标；

（2）行人跟踪。若中心位置变化率小于阈值 1，且余弦相似度大于 0.9，则判定前后帧目标为同一目标。

实现过程中，若存在没有匹配的新目标，则判定为新进入目标，开启新跟踪器进行跟踪。若存在没有匹配的旧目标且连续超过多帧没有匹配，则判定为消失目标，从跟踪器中删除该目标。

值得注意的是，上述跟踪阈值的设置主要是为了避免错误匹配。我们认为，同一目标即使被分段为多个跟踪目标，其最终结果还可以通过行人重识别进行确认，但一旦将两个不同目标作为一个目标跟踪，则会导致结果错误。

（四）人脸识别

利用视频人脸检测具有连续性的特点，能够对前后连续的人脸进行统一编码描述，提高了模糊人脸的描述能力[8]。利用一对多方式进行扩展查询，反向搜索人脸库，提高了查询召回率和准确率。

首先给每张感兴趣的人脸赋予一个 ID，然后将 200 张感兴趣人脸图片进行特征提取，构建人脸信息库，最后将室内视频检测的所有人脸序列与人脸信息库进行匹配，得到视频中人脸序列的 ID 信息。

图5　视频提供的部分人脸图片样例

但我们在实验中发现人脸 ID 库的人脸图片模糊不清（如图 5），与真实场景中人脸图片的差异性较大，提取的人脸特征向量辨识性很差。解决这一问题的算法是，将每一张人脸 ID 库的图片进行输入，反向搜索全部人脸序列，寻找与 ID 库中人脸最相似的视频图片，并最终替换掉原有的人脸图片，构建新的人脸 ID 库。这是一个

一对多与多对一的问题，在实际构建模型时充分考虑这个问题可以有效改善人脸特征表示能力。同时，由于在人脸检测和跟踪过程中人脸往往是一个序列，所以在特征抽取过程中我们通过多张图像的特征进行聚合，能够得到更加有效的人脸特征表示。

人脸特征提取模型的训练和测试过程如图 6 所示，提取特征向量的方法是：计算人脸的 68 个关键点信息，并编码为 128 的一维向量，然后计算不同人脸特征间的欧氏距离。将提取的人脸图像的特征向量与存储的特征库进行搜索匹配，再设定阈值，若相似度超过阈值，则把匹配得到的结果输出。

图 6　人脸行人关联示意图

实验发现，通过一对多反向搜索构建新的人脸 ID 库的方法能够使最终人脸识别的精度提高 27%。

（五）人脸行人关联

根据检测到的人脸和行人的位置关系，若人脸的左上角坐标和右下角坐标均在行人的 bounding box 内，就认为该人脸与行人是匹配的。这样把行人序列和人脸 ID 进行关联，得到行人框 ID 库。

（六）行人再识别

在深度学习模型方面，提出了基于多枝树嵌入的 MTE 模型[7]；在距离测度方面，提出了 GDQS 方法[8]；在多模型融合方面，提出了 HDLF 方法；在 Reid 预测处理方面，我们提出了二重约束的方法。

任务挑战性如下：第一，类内和类间变化。室内外场景的同一个人，由于摄像头角度、分辨率，行人姿态，遮挡等因素，可能导致同一个人的照片出现较大变化[9]。而不同的人在相同场景下也有可能相似，类间距离较小。第二是小样本问题，通过人脸检测和行人检测构造的行人 ID 库样本数量较小，不利于训练和识别。第三是数据标注和泛化能力要求高。最后是对可扩展性和计算实时性要求高。

1. MTE 深度学习模型

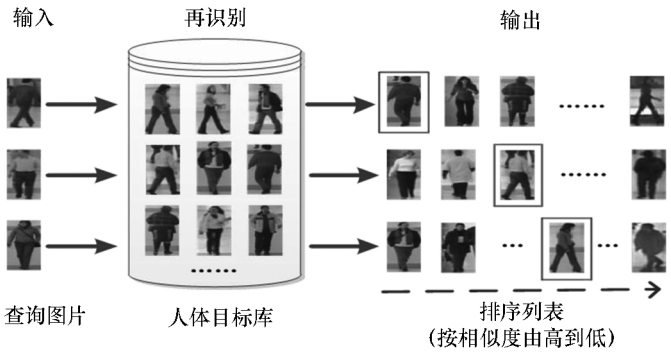

图 7　行人再识别过程图

采用的 MTE 模型如图 8 所示，其中部分缩写为 GAP：Global Average Pooling，全局平均池化；LSR：Label Smooth Regularization，标签平滑规约；FC：Fully – Connected Layer 全连接层。最下面是图像输入，256 × 128 像素的 RGB 彩色图片，输入网络主干。主干网类型多样，如 ResNet 残差网、MobileNet 移动网和 DenseNet 密集网。

MTE 模型使用三类分支：其中 Aux 辅助分枝包含部分中高层特征图接池化、2 个卷积，GAP 和分类器。主枝 Trunk 为 GAP 上面使用多个主枝分类器。Graft 嫁接分枝中，中高层 GAP 后级联，再接 FC 和分类器。损失函数使用三类分支的 Softmax 互熵损失。

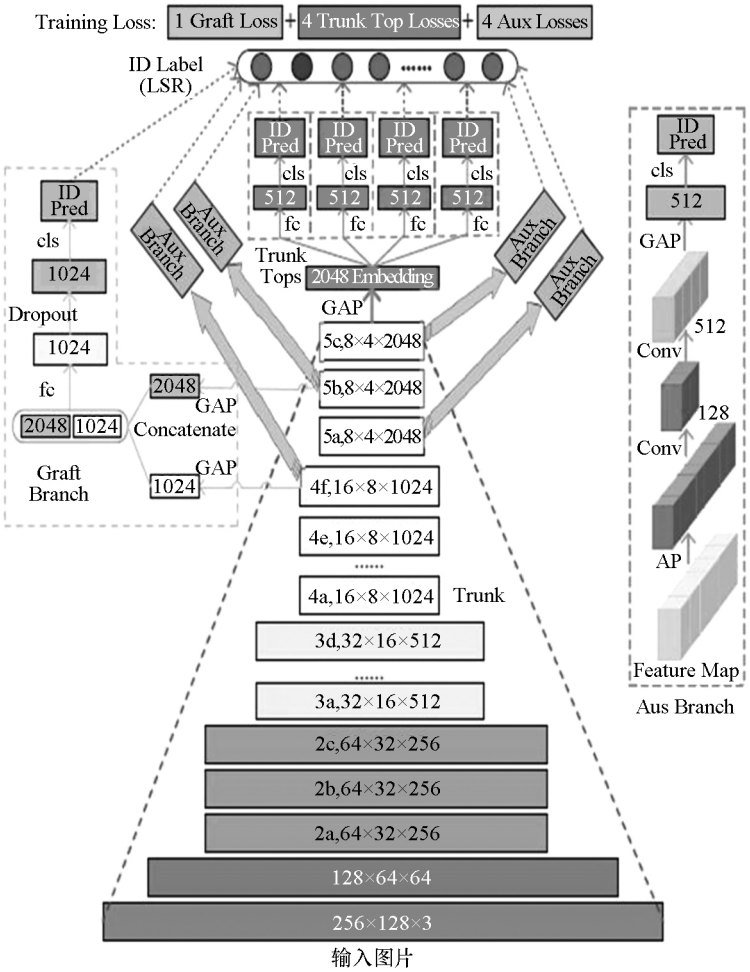

图 8　MTE 模型结构示意图

$$l^{total} = l^{graft} + \sum_{j=1}^{4} l_j^{top} + \sum_{i=1}^{4} w_i l_i^{aux} \quad (1)$$

2. 子空间测度学习方法 GDQS

KISSME[10] 等测度方法需要 PCA 降维步骤，损失区分信息，XQDA 等在特征维度和样本数量均较大时，计算速度较慢。子空间的优势在于可区分性挖掘降维后的计算和存储优势。我们探索了适合 Reid 大规模学习的高效子空间方法。

马氏测度学习

$$d_m(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \quad (2)$$

子空间学习

$$L = L_{d \times d} \quad (3)$$

$$M_{d \times d} = LL^T \quad (4)$$

$$d_L(x_i - x_j) = \| L^T x_i - L^T x_j \|_2 \quad (5)$$

提出的 GDQS 测度方法流程图如图 9，训练集为 X，d 为特征维度，N 为训练样本数。首先经过 ICAR 进行聚合重排，训练样本从大 N 减少到小 n。再经过一个 LSA，分三种情况处理，其中第 2 和第 3 种情况使用了分解。通过 LSA 之后，计算高斯协方差，进行 ERR，然后通过 QDA 和 ML 学习测度，最后得到子空间的转换矩阵 L。其中，ICAR、LSA 和 ERR 是本方法的主要创新点。

标签中心聚合重排步骤如下：首先，将 Probe 和 Gallery 每个行人的图片聚合成多个子集；其次，Probe 和 Gallery 中每个行人计算全局均值特征，其每个子集计算子集均值特征。再次，重排序：按照 Probe 和 Gallery 的行人顺序，先排全局均值，再排子集均值特征。其目的在于后面 QR 分解近似全集时，用前面特征子集可抓取绝大部分数据变化和区分信息。

大规模适配。引入阈值 m = 4000，比较 m 和特征维度 d，样本数量 n 的相对大小，分三种难度处理：

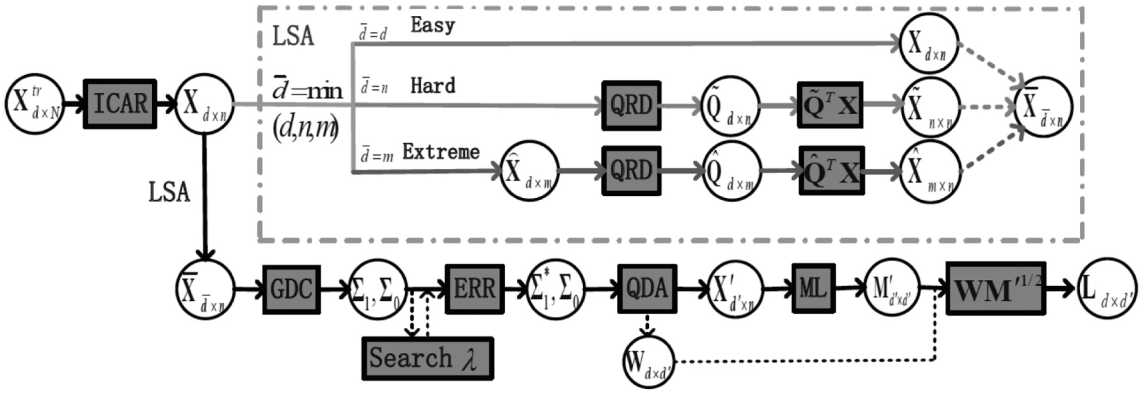

图 9　GDQS 测度方法流程图

$$\bar{X}_{d\times n}=\begin{cases}X_{d\times n},\ \bar{d}=d\ \ (d\leqslant n,\ d\leqslant m)\quad easy\\ \tilde{X}_{n\times n}=\tilde{Q}^{T}X,\ \bar{d}=n\ \ (n<d,\ n<m)\quad hard\\ \hat{X}_{m\times n}=\hat{Q}^{T}X,\ \bar{d}=m\ \ (m<d,\ m<n)\quad extreme\end{cases}\tag{6}$$

其中，简单样本特征维度和样本数量均较小；困难样本特征维度较大，样本数量较小，使用所有训练样本的 QR 分解，可将大值 d 降到小值 n（测度无损，可证）；极难样本特征维度和样本数量均很大，需使用前 m 个均值特征，估计 QR 分解矩阵，将大值 d 或 n 降到阈值 m。

$$x_{ij}=x_i^p-x_j^g\tag{7}$$

$$\sum\nolimits_1=\frac{1}{|\Omega_1|}\sum_{x_{ij}\in\Omega_1}x_{ij}x_{ij}^T\tag{8}$$

$$\sum\nolimits_0=\frac{1}{|\Omega_0|}\sum_{x_{ij}\in\Omega_1}x_{ij}x_{ij}^T\tag{9}$$

ERR。其中，1 代表两个特征来源于同一个人的两种图片，对应计算出的为类内协方差。0 代表不同人的图片，计算出来的为类间协方差。由于类内协方差通常不可逆，采用规约。传统方法使用固定值规约，本文则采用相对规约 ERR，并通过启发式搜索确定最佳规约参数。后续步骤与已有方法类似，本文不再介绍。

$$\sum\nolimits_1^{*}=\sum\nolimits_1+\lambda sI\tag{10}$$

$$s=\sum_{i=1}^{d}\sum\nolimits_1(i,i)/d\tag{11}$$

在 Market1501 图库的实验表明，相比[11]，GDQS 性能相当，但训练时间大幅减少。

3. 多模型多特征融合方法 HDLF

对于单个模型，提取多层特征进行融合的 HDLF 方法使用 MTE。以 ResNet50 为例，单个模型的步骤：训练 MTE，结合 GDQS 确定 HDLF，然后进行 GDQS 得到最终的低维判别子空间。如图 11 所示。

多模型融合：三个模型，每个模型分别提取

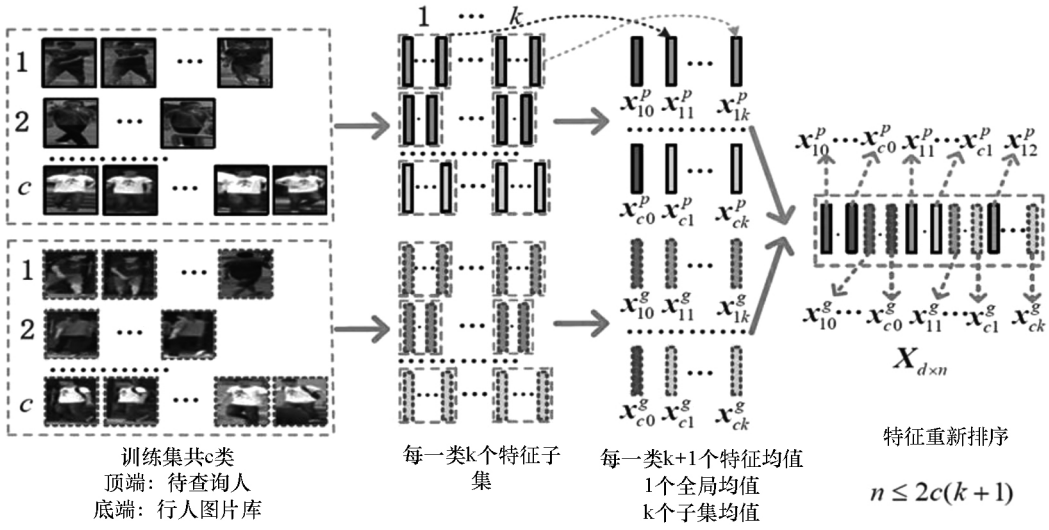

图 10　标签中心聚合重排步骤图

HDLF 多层特征，使用 GDQS 方法映射到低维子空间，多模型的特征加权级联，评估后输出排序列表。

融合方法性能：用三个图库上近四年顶级会议上提出的方法和本文采用方法的性能对比，本文所提方法领先较多。所得结果如图 12 所示。

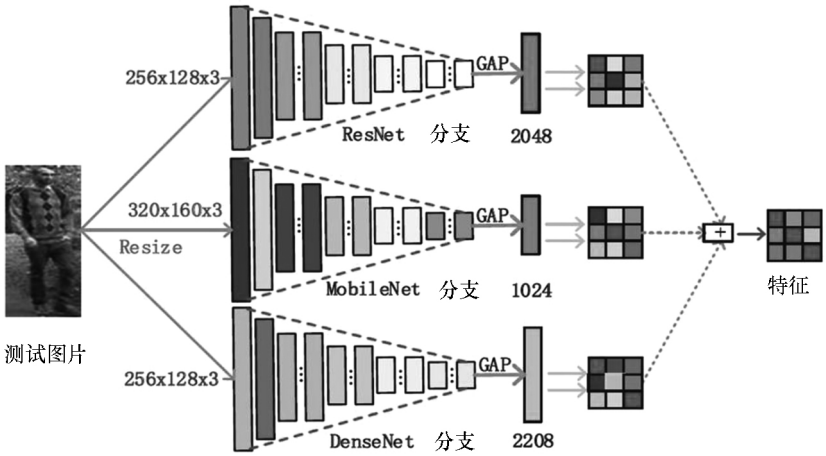

图 11 多模型融合示意图

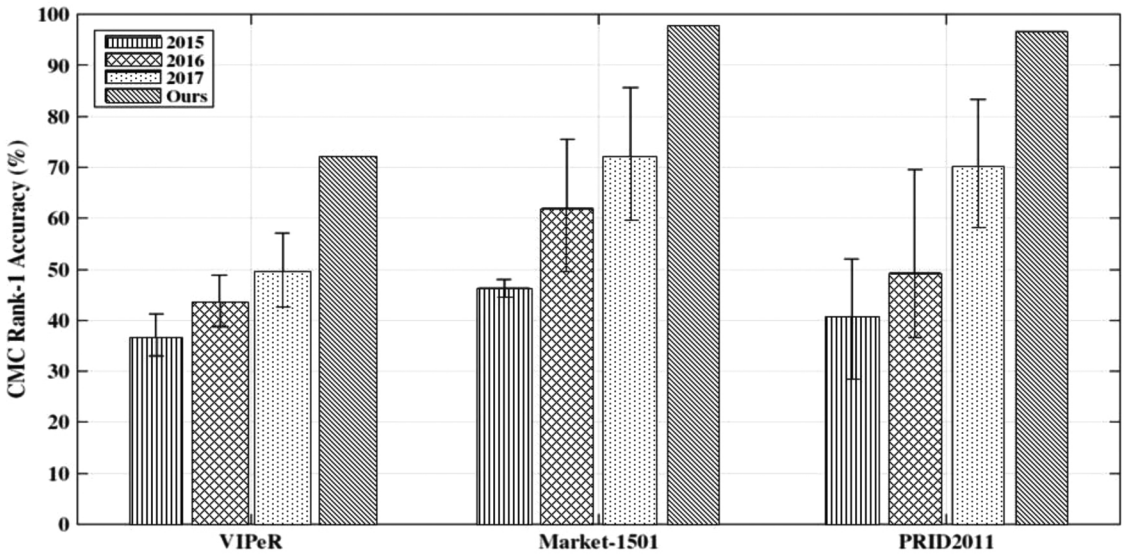

图 12 采用多层特征在 VIPeR、Market－1501、PRID2011 三大图库上的性能比较

4. 对 Reid 预测结果二重约束

经过 Reid 之后输出的结果是每个行人图片预测的 ID 的 rank 排序。对此，我们进行了二重约束。通过 rank 约束和类内距离约束可以调整准确率和召回率。

(1) 对于室内场景的行人图片，如果 rank 前 9 名预测的结果都一致，那么我们认为该张行人图片有可能是我们要寻找的 ID 库中的人，如果 rank 前 9 名中存在不一致的结果，我们认为该张行人图片不是我们要找的 ID 库中的人。同理，对于室外场景，由于在室外任何人之间的差异比室内场景小，需要更严格的约束，因此选择 rank 前 11 的排名来进行约束。

(2) 对第一步判断有可能是 ID 库中出现的行人图片，我们将进行第二重约束，判断是否为真的 ID 库中的人。计算 ID 库中的类内距离 [mean－1×standard deviation，mean＋1×standard deviation]，如果行人图片与 rank1 对应的图的距离在类内距离之内，那么认为该张行人图片的 ID 就是 rank1 预测的 ID，反之则认为是 ID 库以外的图。由于在相同摄像机视角下，人与人的类间差异比较小，所以只采用类内距离进行约束。

(七) 基于时序的后处理

在将行人检测的图片进行行人再识别后，每

一个被查询的 ID 都会输出一个图片序列。事实上同一个人在视频中出现的帧号必然是连续的，我们希望基于帧号对图片序列做一个时序关联。

经过 Reid 之后，将每个 ID 在一个视频中输出的所有图片作为一个目录，提取帧信息构建序列。利用聚类的思想，将整个帧序列分为多个集合，集合与集合之间的间隔为阈值 A，两帧间的间隔超过阈值，那么就认为是不同的两个集合。然后选取整个帧信息中长度最长的集合作为我们预测的该 ID 的行人图片，剔除其他的图片。

我们构建了一个列表用于存贮不同的帧号信息，通过列表的长度来选择最优的集合。

经过测试我们发现阈值选择 70 效果最好（视频 1 秒 25 帧），可以将误识别的图片成功剔除掉。通过该方法我们可以进一步优化 Reid 预测的结果，在训练数据集上的准确率提高了 12.7%。

三、实验

（一）实验环境

实验在 2018 全球人工智能应用大赛多目标跨摄像头跟踪赛题的数据集上测试，实验环境为 ubuntu16.04，高性能计算卡为 4 片 NVIDIA 1080ti 显卡，主要使用的 python 库包含 Opencv3.4.1、Dlib 和 PIL 包等。

人脸检测和人脸识别基于 Face_recognition 包，行人检测和目标追踪模型基于 YOLO V3 的深度学习检测算法，运行框架为 DarkNet，针对行人重新训练，训练测试参数如图 13 所示，行人再识别算法基于 TensorFlow 框架，使用 keras 包。

```
训练数据集：CrowdHuman
实验环境：  ubuntu 16.04, 4 块 1080 Ti 显卡, darknet，python2.7
预训练模型：imagenet 上的 YOLO V3 模型
Python 库：Opencv3.4.1, Dlib, PIL
训练、测试参数：
batch=1 (test)，64 (train)；
subdivisions=1 (test)，16 (train)；
width=416，  height=416；
momentum=0.9；  decay=0.0005；
burn_in=1000; max_batches = 400200;
learning_rate=0.001; policy=steps;
steps=50000,100000,200000; scales=.1,.1,.1
```

图 13　YOLO V3 的训练测试参数

（二）评测指标

为每一个目标赋予一个 ID，通过人脸识别和跨摄像头跟踪把该 ID 赋予视频中每一帧中检测出的行人。如果该行人的检测结果（bounding box）与正确的拥有该 ID 的行人结果高度重合（IoU > 0.5），则该结果是正确的（TP）；否则该结果是错误的（FP）。如果一个正确的拥有该 ID 的行人结果不存在一个被赋予该 ID 的检测结果和它高度重合（IoU > 0.5），则该结果被认为是遗漏的（FN）。

通过两个指标来衡量任务的完成效果，精确度反映预测的精度，召回率反映目标检测的漏检水平。

$$recall = \frac{TP}{TP + FN} \tag{12}$$

$$precision = \frac{TP}{TP + EP} \tag{13}$$

测试算法会逐个计算预测的结果和正确结果中对应目标框的交并比 IOU，若 IOU 的值大于 0.5 则算检测成功，否则失败。统计每个目标的 precision 和 recall 值后计算 F1 - score 作为测评指标，计算公式如下：

$$\text{F1} - score = \frac{2 * precision * recall}{precision + recall} \tag{14}$$

（三）实验结果

我们在多目标跨摄像头赛题排行榜上取得了 0.609 的最佳成绩，最终排名第一。

四、总结与展望

人脸识别在智能城市中已经得到了广泛应用。然而，受限于摄像头的视野和角度，在很多情况下不能保证正面清晰人脸的获得。跨摄像头跟踪技术可以补充人脸识别的局限性，使对行人身份的确认能够得到更大拓展。该技术可以广泛应用于视频监控、智能安保、刑事侦查等领域，通过提供感兴趣目标的人脸图片或行人姿态，搜索该目标出现的时间和地点，并对检测到的目标在视频的帧序列中进行跟踪，分析跟踪的目标轨迹。同时在智能商业方面，通过研究客户运动轨迹和停留时间，获取客户需求，实现客户的私人订制。人工智能即将从“刷脸”跨越到“识人”的新时代。

参考文献

[1] Lin Y，Zheng L，Zheng Z，et al. Improvingperson re - identification by attribute and identity learning [J]. Pattern Recognition，2019：95，151 - 161.

[2] Zheng,Liang,Yang,Yi,Hauptmann,Alexander G. Personre - identification: past,present and future [J]. 2016.

[3] Matsukawa T,Suzuki E. Person re - identification using CNN features learned from combination of attributes [C] // International Conference on Pattern Recognition. IEEE, 2017:2428 - 2433.

[4] Tang X,Du D K,He Z,et al. PyramidBox: Acontext-assisted single shot face detector [J]. 2018.

[5] Redmon J,Farhadi A. YOLOv3: An Incremental Improvement [J]. 2018.

[6] Lin G,Meng Y,Jian Y,et al. Robust,discriminative and comprehensive dictionary learning for face recognition [J]. Pattern Recognition,2018,81:S0031320318301109.

[7] Zeng M, Chang T, Wu Z, et al. Person re - identification by multi - channel feature cascading [C] // Iet International Conference on Wireless. 2014.

[8] Chang T,Zeng M,Wu Z. Personre - identification based on spatiogram descriptor and collaborative representation [J]. IEEE Signal Processing Letters,2015,22(10):1595 - 1599.

[9] Shanshan Jiao, Jiabao Wang, Guyu Hu, et al. Jointattention mechanism for person re - identification [J]. IEEE Access,2019,7,90497 - - 90506.

[10] Zhuang B,Zhu Z,Sourou F A,et al. Person re - identification by the marriage of KISS metric learning and post - rank optimization: KISSPOP [C] // International Conference on Internet Multimedia Computing & Service. 2015.

[11] Liao S, Hu Y, Zhu X, et al. Person re - identification bylocal maximal occurrence representation and metric learning [C] //IEEE Conference on Computer Vision & Pattern Recognition. 2015.

Research on multi-target multi-camera tracking technology

JIAO Shanshan,LI Yunbo,CHEN Jialin,PAN Zhisong

(*Army Engineering University,Nanjing 210000,China*)

Abstract: Multi-target multi-camera tracking refers to finding and correlating the temporal and spatial information of multiple targets in different scenes. The algorithm integrates face detection,pedestrian detection, target tracking,face recognition and person re-identification to form a unified technical framework. Our method can reach the commercial level both in the implementation framework and in the accuracy. We won the second prize in the Global(Nanjing) Artificial Intelligence Application competition.

Key words: multi-target multi-camera tracking; face detection; pedestrian detection; target tracking; face recognition; person re-identification