融合生成对抗网络和姿态估计的视频行人再识别方法

刘一敏 [1] 蒋建国 [1, 2] 齐美彬 [1, 2] 刘 皓 [3] 周华捷 [1]

摘 要 随着国家对社会公共安全的日益重视, 无重叠视域监控系统已大规模的普及. 行人再识别任务通过匹配不同视域摄像机下的行人目标, 在当今环境下显得尤为重要. 由于深度学习依赖大数据解决过拟合的特性, 针对当前视频行人再识别数据量较小和学习特征单一的问题, 我们提出了一种基于视频的改进行人再识别方法, 该方法通过生成对抗网络去生成视频帧序列来增加样本数量和加入了行人关节点的特征信息去提升模型效率. 实验结果表明, 本文提出的改进方法可以有效地提高公开数据集的识别率, 在 PRID2011, iLIDS-VID 数据集上进行实验, Rank 1 分别达到了 80.2 % 和 66.3 %.

关键词 行人再识别, 深度学习, 生成对抗网络, 人体姿态估计

引用格式 刘一敏, 蒋建国, 齐美彬, 刘皓, 周华捷. 融合生成对抗网络和姿态估计的视频行人再识别方法. 自动化学报, 2020, **46**(3): 576−584

DOI 10.16383/j.aas.c180054

Video-based Person Re-identification Method Based on GAN and Pose Estimation

LIU Yi-Min[1] JIANG Jian-Guo[1, 2] QI Mei-Bin[1, 2] LIU Hao[3] ZHOU Hua-Jie[1]

Abstract As the government keeps attaching importance to public security, non-overlapping viewsheds surveillance systems have been deployed widely. It has become especially important to recognize pedestrian target through matching cameras with different viewsheds in nowadays. Deep learning relies on big data to solve overfitting. However, the current video-based person re-identification only has small data volume and homogeneous learning features. To solve this, we put forward a method to improve person re-identification based on the video. This method can increase the sample quantity by generating video frame sequence through generative adversarial network. It also adds the feature information of the pedestrian joints, which can improve the model efficiency. The experiment result shows that the modified method discussed in this paper can improve the recognition rate of public datasets effectively. In the experiments on PRID2011 and iLIDS-VID, Rank 1 attained 80.2 % and 66.3 %, respectively.

Key words Person re-identification, deep learning, generative adversarial network (GAN), human pose estimation

Citation Liu Yi-Min, Jiang Jian-Guo, Qi Mei-Bin, Liu Hao, Zhou Hua-Jie. Video-based person re-identification method based on GAN and pose estimation. *Acta Automatica Sinica*, 2020, **46**(3): 576−584

行人再识别是指在无重叠视域监控系统中, 检测和匹配在不同视域摄像头下两个行人是否为一个人的任务. 广泛应用于公安实时监控区域、捉捕嫌疑人等任务当中, 对维护国家的治安、提高公安办案效率有着重要的意义. 但是由于不同视域下的摄像头之间的行人存在分辨率、视角、光照条件、遮挡、背景干扰、行人姿态以及摄像头成像质量的差异, 导致相同的人在不同视域摄像头下存在很大的外观差异, 会给行人再识别带来很大挑战. 深度学习在计算机视觉上取得的巨大成功, 使得在行人再识别领域的研究也日益增加, 卷积神经网络 (Convolutional neural network, CNN) 作为特征提取器用来自主学习特征, 文献 [1] 提出了利用 Siamese 网络进行有监督的学习, 来匹配和区分行人对. 文献 [2] 将长短时记忆网络 (Long short-term memory, LSTM) 引入 Siamese 网络, 对行人图像进行分割, 用 LSTM 依次捕捉各区域之间的空间关系, 增强网络的判别能力. Liu 等[3] 提出一个端到端的基于比较性注意力网络, 使用 LSTM 来循环生成局部注意力的特征, 提取到更多局部的辨别性信息, 有效提高行人再识别算法性能. 由于实际环境条件的需求, 基于视频的行人再识别工作越来越得到关注. Wang 等[4] 提出的一个区分视频片段的方法 (Discriminative video fragments selection and ranking, DVR) 框架用于行人再识别, 使用判别性的时空特征选择来自动发

收稿日期 2018-01-22 录用日期 2018-07-02
Manuscript received January 22, 2018; accepted July 2, 2018
国家自然科学基金 (61371155, 61771180), 安徽省重点研究与开发项目 (1704d0802183) 资助
Supported by National Natural Science Foundation of China (61371155, 61771180) and Anhui Province Key Research and Development Projects (1704d0802183)
本文责任编委 刘青山
Recommended by Associate Editor LIU Qing-Shan
1. 合肥工业大学计算机与信息学院 合肥 230009 2. 工业安全与应急技术安徽省重点实验室 合肥 230009 3. 腾讯优图实验室 合肥 230009
1. School of Computer and Information, Hefei University of Technology, Hefei 230009 2. Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei 230009 3. YouTu Laboratory, Tencent, Hefei 230009

现和利用最可靠的视频片段. 文献 [5] 使用自适应的 Fisher 判别分析来解决视频行人再识别问题. 文献 [6] 利用视频三元组在训练视频中学习视频内距离度量和视频间距离度量, 从而提高识别的准确度. 文献 [7] 建立了时空人体动作模型, 可以为视频中的行人构建很好的时空表示. McLaughlin 等[8] 提出了卷积神经网络捕捉图像特征, 循环神经网络 (Recurrent neural networks, RNN) 捕捉帧与帧的时空关系特征来配合 Siamese 网络, 取得了很好的实验结果. 文献 [9] 使用 LSTM 网络以循环的方式来聚合每帧的行人特征. 文献 [10] 采用端到端的双流网络, 提取了行人图片特征和运动上下文信息来提高识别率. 文献 [11] 在文献 [8] 的基础上分别添加了 CNN 和 RNN 上的注意力机制, 使其在时空上有重点的关注某些重要特征, 提升了识别效率.

2014 年, Goodfellow 等[12] 提出了生成对抗网络 (Generative adversarial networks, GAN). 生成对抗网络来源于二人零和博弈思想的启发, 结构分为两个部分, 生成网络 (Generator network, G) 和判别网络 (Discriminator network, D), 生成网络主要是通过捕捉训练集上的数据, 来产生新的样本, 而判别网络则是判断样本是否是生成网络生成的还是来自原训练集. 两种网络相互竞争, 最后判别器将无法区分训练数据分布和生成数据分布, 生成的图片满足原训练集的图片分布, 则完成训练, 实验表明在伪造图片方面, GAN 网络具有很强的优势. 文献 [12] 的提出使得越来越多的基于生成对抗网络的研究被提出, 文献 [13] 将 GAN 扩展到了 CNN 的领域, 使得 GAN 的训练更加稳定和可控. 文献 [14] 在原始 GAN 的基础上加入了监督的信息, 使得生成的图片向着标签方向生成. 文献 [15] 采用序列化的思想, 结合图像拉普拉斯金字塔实现序列化的生成, 减少 GAN 每次学习的内容和难度, 图像质量得到提升. 为了解决 GAN 模型中存在训练困难、生成器和判别器的损失函数无法指示训练进程、生成样本缺乏多样性等问题, 文献 [16] 采用 Earth-Mover 距离替代了 JS (Jensen-Shannon) 散度来衡量距离, 在近似最优判别器下优化生成器缩小 Wasserstein 距离, 拉近生成分布与真实分布.

人体姿态估计是通过给定图像来确定图像中各个人体部位的位置的过程, 用来分析人体动作和行为. 传统的方法分为基于人体特征[17−19] 和基于模型的方法[20]. 文献 [17] 使用形状上下文特征作为人体的外观特征, 利用正则最小二乘法和支持向量机来进行回归. 文献 [18] 同样用形状上下文作为行人特征, 通过距离度量来判断图像间的相似度, 用来分析行人姿态估计. 文献 [19] 用 Gist 特征来作为人体特征, 通过进行非线性近邻元分析图像相似度来确定人体手和头部的姿态. 由于人体非刚体的特性, 基于模型的方法被提出, 文献 [20] 设想人体不同部位是相互独立存在的, 并对不同部位添加了约束, 构造了人体的树形图结构模型作为姿态估计的模型. 基于深度学习的姿态估计方法近几年也被提出, 并取得了不错的成绩. 文献 [21] 用卷积神经网络来提取人体特征并计算关节点的分布, 结合现有的人体模型进行姿态估计. 文献 [22] 利用 CNN 来估计人体姿态, 融合空间信息和光流信息, 并且用热力图 (Heatmap) 来取代关节点的坐标, 提高了关节点检测的鲁棒性.

目前在行人再识别课题研究中, 大量行人图像重构和生成的行人再识别方法被提出. 高质量的行人图片和大量的标签样本有助于网络性能的提高. 文献 [23] 提出了一种半耦合低秩判别字典学习 (Semi-coupled low-rank discriminant dictionary learning, SLD^{2}L) 方法, 将超分辨率重建引入行人再识别研究. Zheng 等[24] 将生成对抗网络应用于行人再识别方向, 通过生成对抗网络生成符合行人特征分布的图像, 并平滑行人标签, 提高了匹配效果. Qian 等[25] 利用生成对抗网络生成出不同姿势的行人图片, 解决了应用场景下行人姿势的干扰, 提升了算法识别效果. 为了解决在一个数据集上训练的网络不能应用在另外一个数据集的问题, 文献 [26] 设计了 SPGAN (Similarity preserving cycle-consistent generative adversarial network), 可以无监督地将标记图像从源域转换为目标域.

我们相信在实验中添加带有标签的样本可以提升网络性能, 提高识别效果. 因此本文提出了利用生成对抗网络来生成带有标签的视频样本通过增加样本数来提高网络能力, 同时在输入中增加了行人关节点的一维特征. 实验显示, 基于文献 [11] 的方法, 相较于其他行人再识别相关方法, 在 PRID2011 和 iLIDS-VID 基于视频的行人再识别数据集上实验, 行人匹配率得到了显著的提升.

本文的其余章节组织安排如下. 第 1 节介绍本文提出的融合生成对抗网络和姿态估计的视频行人再识别方法; 第 2 节介绍本文算法在视频行人再识别公共数据集上的实验; 第 3 节总结全文以及展望.

1 融合生成对抗网络和姿态估计的视频行人再识别方法

1.1 通过生成对抗网络预测帧以增加样本

文献 [27] 采用了一种图像多尺度结构 (见图 1) 结合生成对抗网络对抗训练方法, 并设计了图像梯度差损失函数来保证生成帧的清晰度. 为了避免采样和池化所带来的分辨率上的损失, 采用了拉普拉

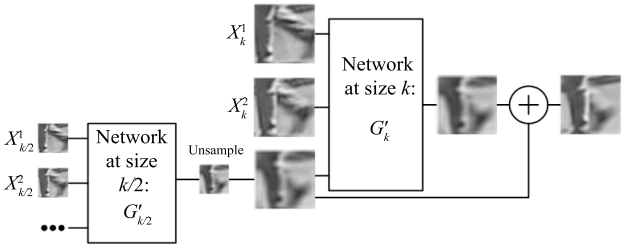


图 1 多尺度结构

Fig. 1 Multi-scale architecture

斯金字塔的结构, 通过 4 次升采样不断的逼近真实样本. 拉普拉斯金字塔中各层生成网络生成的预测图像可以表示为

$$\hat{Y}_k = G_k(X) = u_k(\hat{Y}_{k-1}) + G_k^{'}(X_k, u_k(\hat{Y}_{k-1})) \quad (1)$$

其中, k 下标表示不同的输入图片尺寸, k 取值从 1 到 4, 分别代表输入尺寸为 4×4, 8×8, 16×16 和 32×32, u_k 表示图像升采样到 k 尺寸大小的图像, $G_k^{'}$ 表示输入图像大小符合 k 尺寸的生成模型, $\hat{Y}_k$ 表示生成模型生成的符合 k 尺寸的预测图像. 判别模型 D 输入一系列的图像帧, 用来训练分辨序列的最后一帧是真实的图像还是 G 生成的图像. 同样为了满足生成模型中不同尺寸的变化, D 是具有单个标量输出的多尺度卷积网络. G, D 交替训练, 使用随机梯度下降算法 (Stochastic gradient descent, SGD) 来最小化损失函数. 判别模型的损失函数 $L_{\rm adv}^D$ 表示为

$$L_{\rm adv}^D(X,Y) = \sum_{k=1}^{N_{\rm scales}} \big[L_{\rm bce}(D_k(X_k, Y_k), 1) + L_{\rm bce}(D_k(X_k, G_k(X)), 0)\big] \quad (2)$$

其中, (X, Y) 是来自数据集的样本. X 是 m 个帧的序列, Y 是 X 下一帧的图像. 对每个输入的尺寸 k, 固定生成模型 (G), 对 D 进行 SGD 迭代, 训练判别模型可以将真实输入分类到 1 类, 伪造输入分类到 0 类. $L_{\rm bce}$ 表示的是二元交叉熵损失函数, 表示为

$$L_{\rm bce}(Y, \hat{Y}) = -\sum_i \big[\hat{Y}_i \lg(Y_i) + (1-\hat{Y}_i)\lg(1-Y_i)\big] \quad (3)$$

生成模型的损失函数 $L_{\rm adv}^G$ 表示为

$$L_{\rm adv}^G(X,Y) = \sum_{k=1}^{N_{\rm scales}} L_{\rm bce}(D_k(X_k, G_k(X_k)), 1) \quad (4)$$

其中, (X, Y) 同式 (2). 固定判别模型 (D), 使用 SGD 对 G_k 进行迭代以最小化损失函数, 在生成模型中, 最小化损失函数的目的就是生成的图片尽可能地去迷惑判别模型, 使判别模型不能正确地区分真实图片和生成图片. 但是在实验中, 这种损失函数的构建会导致系统不稳定, 判别网络会生成混乱的样本来欺骗判别网络, 因此设计联合损失函数 $L(X, Y)$, 在生成模型损失函的基础上加入 L_p 损失函数以及图像梯度差分损失 $L_{\rm gdl}$ 去联合优化生成网络, 表示为

$$L(X,Y) = \lambda_{\rm adv} L_{\rm adv}^G(X,Y) + \lambda_{l_p} L_p(X,Y) + \lambda_{\rm gdl} L_{\rm gdl}(X,Y) \quad (5)$$

其中, $\lambda_{\rm adv}$, λ_{l_p} 和 $\lambda_{\rm gdl}$ 分别为 $L_{\rm adv}^G$, L_p 和 $L_{\rm gdl}$ 损失函数的权重, $L_p(X, Y)$ 通过最小化与真实图片的距离来优化模型, p 的值可以是 1 或者是 2, L_p 表示为

$$L_p(X,Y) = l_p(G(X), Y) = \|G(X) - Y\|_p^p \quad (6)$$

$L_{\rm gdl}$ 表示为

$$L_{\rm gdl}(X,Y) = L_{\rm gdl}(\hat{Y}, Y) = \sum_{i,j} \Big[\Big||Y_{i,j} - Y_{i-1,j}| - |\hat{Y}_{i,j} - \hat{Y}_{i-1,j}|\Big|^a + \Big||Y_{i,j-1} - Y_{i,j}| - |\hat{Y}_{i,j-1} - \hat{Y}_{i,j}|\Big|^a\Big] \quad (7)$$

式 (7) 是计算生成图片和真实图片像素之间梯度差异, 其中, a 是大于或者等于 1 的整数, $|\cdot|$ 表示绝对值函数. 加入梯度差分损失函数作为惩罚项, 目的是使生成图像更加锐利.

实验中通过生成对抗网络递归生成所有行人视频序列下一帧的图像, 并利用循环递归来生成后 N 帧的图像, 并给每帧图像打上该行人身份标签. 图 2 中, 每行的前 5 帧图像是视频行人再识别数据集上的连续图像, 而每行的后 5 帧是通过前 5 帧的输入由生成对抗网络生成出来的连续图像. 可以看出, 通过训练, 生成的后 5 帧图像保持了行人正常的步伐轨迹, 同时有较好的图像效果. 通过后面实验论证, N 值取 5 帧时算法性能最优, 本文实验统一设 $N = 5$. 实验中生成图像进行迭代的过程中误差的不断叠

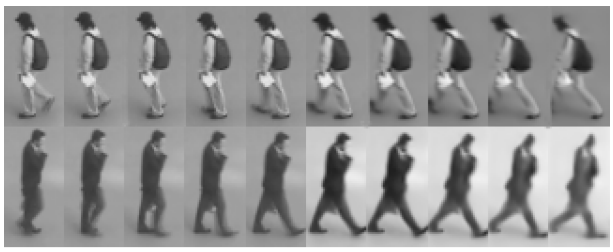


图 2 生成对抗网络生成的视频帧序列 (后 5 帧)

Fig. 2 A sequence of video frames generated by GAN (last five frames)

加, 导致迭代 5 帧之后的图片质量难以保证, 会影响行人正确的识别和分类.

1.2　行人关节点的检测

目前视频行人再识别的特征主要是神经网络通过原图和光流图来提取特征, 原图可以提取到行人空间上的深度特征, 光流是图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧与当前帧之间存在的对应关系的时间特征. 我们认为加入行人关节点的一维有效特征会帮助网络正确分类, 因为通过视频中行人关节点的变化情况, 可以提取到关节点运动的时间信息帮助网络分类, 也会让网络关注关节点周围的信息, 避免了背景所带来的影响.

文献 [28] 在文献 [29] 的基础上设计了卷积姿态机 (Convolutional pose machines, CPM) 算法, 算法利用了全卷积的网络学习图像特征和图像间的空间信息来解决姿态估计任务. 在任务中设关节点数目为 p, 关节点的坐标用 Y_p 来表示, $\mathcal{Z}$ 表示所有关节点坐标的集合, 存在 $\forall Y_p \in \mathcal{Z}$, 算法目的是预测出 p 个点的坐标 Y_p.

图 3 (a) 中使用了深度卷积网络来进行局部预测, 使用了五个常规卷积层和两个 $1*1$ 的卷积层构成的网络结构, 并在网络的最后将局部图回归到一个 $p+1$ 大小的输出矢量, 表示了图像区域每个关节点的得分信息. 图 3 (b) 中展示了后续网络的结构, 网络不仅仅会得到原图的特征也会加入上一阶段输出的特征来表示图像上下文之间的联系. 并且通过大尺度的卷积 $(11*11)$ 增加检测图像感受野的尺度, 使得关节点的检测既能确保精度, 又考虑了各个关节点之间的距离关系. 作为一种多阶段的网络结构, 为了防止网络训练梯度传播过程中发生的梯度消失现象, 网络会在每个阶段内添加监督信息, 有效地加强反向传播能力, 避免发生梯度消失. 每个阶段优化任务表示为

$$f_t = \sum_{p=1}^{P+1} \sum_{z \in \mathcal{Z}} \| b_t^p(z) - b_*^p(z) \|_2^2 \tag{8}$$

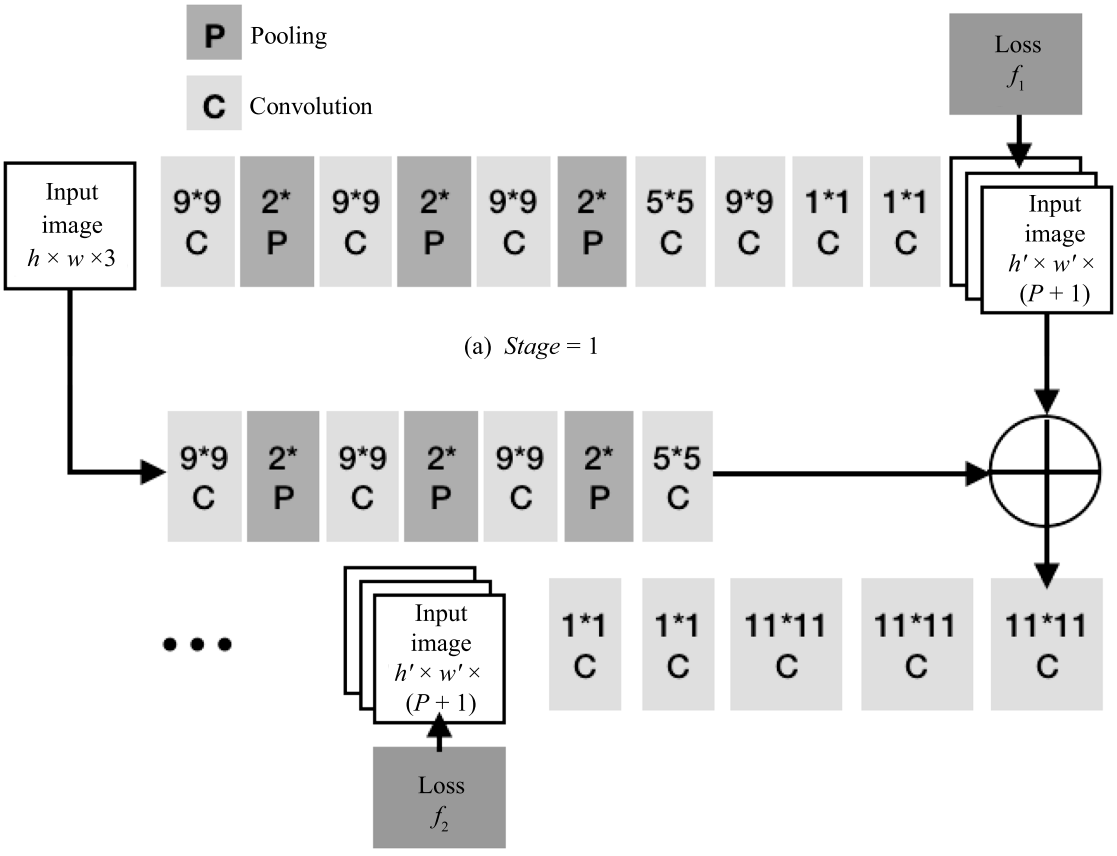


图 3　CPM 算法的网络结构

Fig. 3　Structure of CPM algorithm

其中, b_t^p 表示 CPM 网络结构每个阶段 t 输出的关节点 p 的置信图, b_*^p 表示原图中对图像标注中的真实关节点 p 做高斯分布产生. 而损失函数在每个阶段 t 的优化函数是每个像素间的平方误差.

整个 CPM 网络优化函数表示如式 (9), 即每个阶段 t 优化函数的叠加. 网络使用标准随机梯度下降来优化网络性能.

$$\mathcal{F} = \sum_{t=1}^{T} f(t) \tag{9}$$

文献 [30] 基于 CPM 算法设计了一种复杂场景下实时多人姿态检测的方法, 解决了运算效率低和多人分类不准的问题, 实验在 MPII 和 MSCOCO Keypoint 数据集上都展现了很好的效果. 网络结构主要由部位置信图 (Part confidence maps, PCM) 和部位亲和域 (Part affinity fields, PAF) 组成, 利用 PCM 来回归关节点和 PAF 回归出两个关节点之间的热图, 最后通过图论中偶匹配的方式, 连接两个关节点, 获取多人的姿态.

目前的行人再识别数据集中还没有关节点的标签, 利用文献 [30] 的算法在 MSCOCO Keypoint 数据集上的训练, 可以获得多达 18 个关节点的行人关节点特征. 如图 4 所示, 上列是行人连续运动的视频帧序列, 下列则是根据上列的图像来可视化的人体关节点的一维图像信息, 可以看出很好的检测并可视化出行人 18 个关节点信息, 描述出行人动作和姿态的特征.

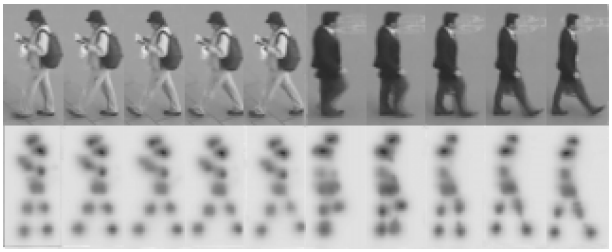


图 4 CPM 算法检测到的行人关节点特征

Fig. 4 Pedestrian keypoint features detected by CPM algorithm

1.3 本文算法的具体步骤

步骤 1. 输入视频行人再识别的数据集, 利用多

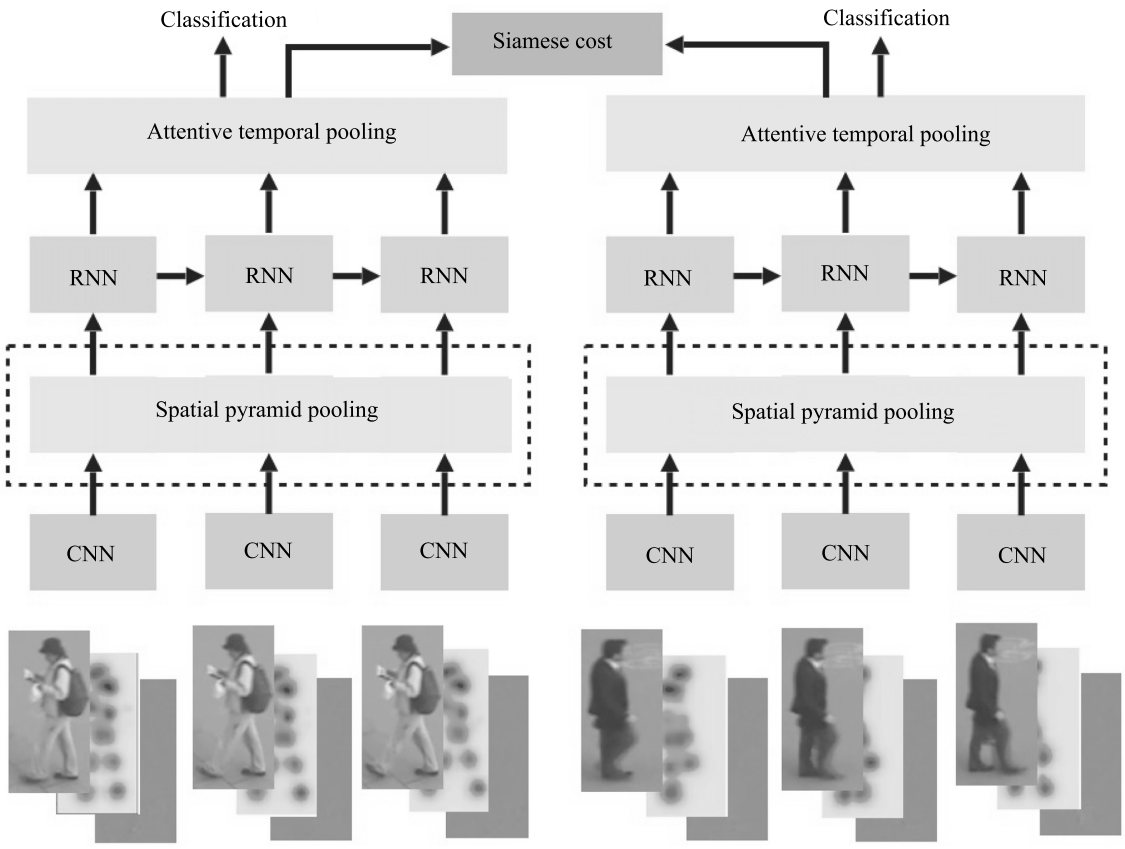


图 5 融合生成对抗网络和姿态估计算法网络结构

Fig. 5 The structure of integration of GAN and pose estimation algorithm

尺度生成对抗网络去预测下一帧的图像, 用第 1.1 节中式 (5) 作为优化任务, 循环递归生成 N 帧图像, 生成的图像将打上该序列中行人身份标签.

步骤 2. 利用第 1.2 节中图 3 的网络模型联合式 (8) 和式 (9) 的优化任务对扩充的视频行人再识别数据集进行人体姿态估计, 获得数据集中不同行人关节点的一维信息特征.

步骤 3. 利用步骤 1 扩充的行人再识别数据集和步骤 2 生成的数据集中人体关节点的信息, 与 YUV 颜色三通道和光流两通道组成 6 通道图像, 使用时–空注意力网络 (Attentive spatial-temporal pooling networks, ASTPN) 进行分类, 获取实验结果.

2　实验测试与结果

本节首先介绍了行人再识别研究中算法性能的评测准则和实验中所使用的数据集, 其次介绍了本文算法在不同公共实验数据集上的实验结果, 并与已有的行人再识别算法实验性能进行了比较, 分析了两个创新部分分别对算法性能产生的影响. 最后本节分析了由生成对抗网络递归产生的帧数对算法性能的影响, 得出的结论使算法性能达到最优. 文中的实验通过 Torch 和 TensorFlow 框架实现, 硬件采用搭载 i5-4590 (3.30 GHz) 和 NVIDIA GTX-980TI (4 GB) 的个人电脑.

2.1　测试数据和算法性能的评测准则

本文在 PRID2011 和 iLIDS-VID 视频行人再识别数据集上进行实验与测试. 基于文献 [11] 的描述, 将视频行人再识别数据集上的行人对随机分为两个大小相同的子集, 一个为训练集. 另一个为测试集. 由于在视频行人再识别数据集上查询集 (Probe) 和行人图像集 (Gallery) 序列是长度不固定的, 所以在训练过程中每个训练过程随机地选择 $k = 16$ 个连续帧的子序列. 在测试过程中, 我们将第 1 台相机视为查询集, 将第 2 台相机视为行人图像集. 并给同一个行人来自摄像机 1 的子序列和来自摄像机 2 的子序列打上正确行人对的标签, 而负行人对标签来自不同人在摄像机 1 和摄像机 2 的子序列. 正标签对和负标签对经过图像增强会被连续地送入 ASTPN 网络, 使模型能够区分正确匹配和错误匹配. 实验采用累积匹配特征曲线 (Cumulative match characteristics, CMC) 来评价算法的性能. CMC 在实验中表示在不同摄像机下行人匹配的概率. 实验中给定一个查询的数据集和行人图像数据集 (摄像机 1 和摄像机 2), 累积匹配特征曲线描述的是在行人图像数据集中搜索待查询数据集中的行人, 前 r 个搜索结果中显示找出待查询人查询到的比率. 在日常生活中, 我们人类衡量匹配的准确率往往是第一匹配率 (Rank 1), 因此我们在实验中同样将第一匹配率放在衡量算法性能标注的首要位置. 每组实验将训练集和测试集过程重复 10 次, 取 10 次的平均值作为该组实验的结果.

2.2　在不同数据集上的实验结果

2.2.1　PRID2011 数据集

PRID2011 数据集中有 385 个行人视频序列来自摄像机 A, 有 749 个行人视频序列来自摄像机 B. 但是只有 200 个人同时出现在两台摄像机中, 每个行人序列的长度在 5 到 675 个图像帧不等, 平均行人视频帧数为 100 帧, 整个数据集总共有 24 541 张图片, 图片尺寸为 128 像素 × 64 像素大小.

表 1 给出了本文改进方法针对 PRID2011 数据集相对于其他算法累积特征匹配曲线的对比, 可以看出本文改进方法相较于 ASTPN 和其他基于视频的行人再识别方法来说, 可以有效地提升 CMC 中 Rank 1 的识别率, 对 Rank 5 的识别率也有提升. 对比 ASTPN 算法, 加入更多样本和特征的信息会让 Rank 1 提升了 3.2 %.

表 1　不同算法在 PRID2011 数据集上的识别率 (%)
Table 1　Matching rates of different methods on the PRID2011 dataset (%)

方法	Rank 1	Rank 5	Rank 10	Rank 20
AFDA[5]	43.0	72.7	84.6	91.9
VR[4]	41.8	64.5	77.5	89.4
STA[7]	64.1	87.3	89.9	92.0
RFA[9]	64.1	85.8	93.7	98.4
RNN-CNN[8]	70.0	90.0	95.0	97.0
ASTPN[11]	77.0	95.0	99.0	99.0
本文方法	**80.2**	**96.0**	**99.1**	**99.2**

从表 2 可以看出, 在 PRID2011 数据集上通过生成对抗网络生成更多带标签的样本和添加了行人关节点的特征信息, 都对行人再识别的 Rank 1 有所提高, 分别提升 2.2 % 和 1.6 %.

2.2.2　iLIDS-VID 数据集

iLIDS-VID 数据集是从机场的两个无重叠视域的摄像机中捕捉到的行人轨迹图像. 数据集中包含了 300 个不同行人的 600 个图像视频序列, 相同行人的两个视频序列分别采集于摄像机 1 和摄像机 2. 在数据集中每个行人视频序列的长度在 23 到 192 个图像帧不等, 每个视频序列的平均帧数为 73 帧, 数据集中总共有 42 495 张图片. 数据集中包含了人与人之间的服装相似性, 不同摄像机光线和视角的

表 2 不同算法在 PRID2011 数据集上对识别率的影响 (%)
Table 2 The influence of different methods on matching rates based on PRID2011 dataset (%)

方法	Rank 1	Rank 5	Rank 10	Rank 20
ASTPN	77.0	95.0	99.0	99.0
ASTPN+GAN	79.2	95.3	99.2	99.2
ASTPN+KeyPoint	78.6	95.1	99.1	99.1
本文方法	**80.2**	**96.0**	**99.1**	**99.2**

变化, 复杂的背景以及存在的遮挡等问题, 给研究人员带来了很大的挑战.

由表 3 可知, 本文改进方法对 iLIDS-VID 数据集同样性能上有所提高, 将 Rank 1 提高到 66.3%. 相较于 ASTPN 有效提高 4.3%, Rank 5 也提升了 2.4%.

表 3 不同算法在 iLIDS-VID 数据集上的识别率 (%)
Table 3 Matching rates of different methods on the iLIDS-VID dataset (%)

方法	Rank 1	Rank 5	Rank 10	Rank 20
AFDA[5]	37.5	62.7	73.0	81.8
VR[4]	34.5	56.7	67.5	77.5
STA[7]	44.3	71.7	83.7	91.7
RFA[9]	49.3	76.8	85.3	90.1
RNN-CNN [8]	58.0	84.0	91.0	96.0
ASTPN[11]	62.0	86.0	94.0	98.0
本文方法	**66.3**	**88.4**	**96.2**	**98.1**

从表 4 可以看出, 在 iLIDS-VID 数据集上, 改进的两种方法都对算法性能有所提升, 分别将 Rank 1 提升 2.4 % 和 2.5 %, Rank 5 提升 1.5 % 和 1.5 %.

表 4 不同算法在 iLIDS-VID 数据集上对识别率的影响 (%)
Table 4 The influence of different methods on matching rates based on iLIDS-VID dataset (%)

方法	Rank 1	Rank 5	Rank 10	Rank 20
ASTPN	62.0	86.0	94.0	98.0
ASTPN+GAN	64.4	87.5	95.1	98.0
ASTPN+KeyPoint	64.5	87.5	96.1	98.1
本文方法	**66.3**	**88.4**	**96.2**	**98.1**

2.3 生成对抗网络递归生成 N 张图片对方法性能的比较

实验发现, 通过生成对抗网络对视频行人再识别的生成中, 每个行人序列所生成的样本数对行人再识别的累积特征匹配曲线有影响。PRID2011 数据集中, 有 200 对有效视频行人再识别序列, N 是每个序列生成对抗网络循环递归产生图片的数量, 数据集会产生 $N \times 400$ 张图片. 同理, iLIDS-VID 数据集中有 300 对有效视频行人再识别序列, 故数据集会产生 $N \times 600$ 张图片. 表 5 和表 6 分别给出了本文改进方法和 N 的大小对 PRID2011 和 iLIDS-VID 数据集算法性能影响, 并都加入了行人关节点的一维信息.

表 5 每个行人轨迹递归生成的图片张数 N 对 PRID2011 数据集上识别率的影响 (%)
Table 5 The influence of the number N of pictures generated recursively by each pedestrian trace on matching rate based on PRID2011 dataset (%)

N	Rank 1	Rank 5	Rank 10	Rank 20
1	77.4	95.2	99.0	99.0
3	78.6	95.6	99.1	99.1
5	**80.2**	**96.0**	**99.2**	**99.3**
7	80.1	**96.0**	**99.2**	99.2
9	77.6	95.5	98.7	99.1
11	76.7	95.4	98.2	99.0

表 6 每个行人轨迹递归生成的图片张数 N 对 iLIDS-VID 数据集上识别率的影响 (%)
Table 6 The influence of the number N of pictures generated recursively by each pedestrian trace on matching rate based on iLIDS-VID dataset (%)

N	Rank 1	Rank 5	Rank 10	Rank 20
1	61.9	86.7	94.2	97.8
3	63.1	87.5	95.6	98.0
5	**66.3**	**88.4**	**96.2**	**98.1**
7	66.0	**88.4**	96.0	**98.1**
9	64.8	87.9	94.3	97.9
11	64.6	86.6	94.1	97.6

从表 5 和表 6 可知, 当 $N = 1$ 到 $N = 5$ 之间, 算法性能会逐渐上升, 在 $N = 5$ 和 $N = 7$ 之间算法性能达到最优, 而在 $N > 7$ 之后, 算法性能逐渐下降, 考虑到算法效率结合最优结果, 本实验会将 N 值设为 5. 通过实验验证, 每个行人视频序列生成的 N 并不是越多越好, 由于生成对抗网络递归中误差的不断叠加, 在后续产生的图像中行人特征会变差, 干扰了网络特征提取和正确分类.

3 结束语

现有的基于视频的行人再识别方法, 主要从设计提取鲁棒的时间空间特征深度网络和设计泛化性

能好的分类器这两个方向入手. 但是现有视频行人再识别数据集一般规模较小, 没有充分的监督样本来进行网络的训练和优化, 容易使深度网络发生过拟合的现象, 泛化性不高. 对此, 本文提出融合生成对抗网络和姿态估计的视频行人再识别方法. 该方法首先利用多尺度生成对抗网络去生成行人带标签的视频序列, 生成的视频序列无论行人外貌和运动轨迹都接近于真实视频序列, 有效扩大了视频行人再识别数据集的数据量; 其次加入了视频行人关节点的信息, 增加了行人的姿态特征; 最后利用了时−空注意力网络对处理过的数据集进行分类, 判断样本之间的相似性. 两个视频行人再识别数据集上的实验相较于其他改进算法也取得了较好的实验结果. 针对生成的图像序列数目有限的问题, 下一步需要解决在多次迭代中图像误差逐渐增大的问题, 产生更多连续高质量图像来扩增数据集.

References

1 Yi D, Lei Z, Liao S C, Li S Z. Deep metric learning for person re-identification. In: Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014. 34−39

2 Varior R R, Shuai B, Lu J W, Xu D, Wang G. A Siamese long short-term memory architecture for human re-identification. In: Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 135−153

3 Liu H, Feng J S, Qi M B, Jiang J G, Yan S C. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017, **26**(7): 3492−3506

4 Wang T Q, Gong S G, Zhu X T, Wang S J. Person re-identification by video ranking. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 688−703

5 Li Y, Wu Z Y, Karanam S, Radke R J. Multi-shot human re-identification using adaptive fisher discriminant analysis. In: Proceedings of the 2015 British Machine Vision Conference (BMVC). Swansea, UK: BMVA Press, 2015. 73.1−73.12

6 Zhu X K, Jing X Y, Wu F, Feng H. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI). New York, USA: ACM, 2016. 3552−3558

7 Liu K, Ma B P, Zhang W, Huang R. A spatio-temporal appearance representation for video-based pedestrian re-identification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 3810−3818

8 McLaughlin N, del Rincon J M, Miller P. Recurrent convolutional network for video-based person re-identification. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1325−1334

9 Yan Y C, Ni B B, Song Z C, Ma C, Yan Y, Yang X K. Person re-identification via recurrent feature aggregation. In: Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 701−716

10 Liu H, Jie Z Q, Jayashree K, Qi M B, Jiang J G, Yan S C, et al. Video-based person re-identification with accumulative motion context. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017, DOI: 10.1109/TCSVT.2017.2715499

11 Xu S J, Cheng Y, Gu K, Yang Y, Chang S Y, Zhou P. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4743−4752

12 Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS). Montreal, Canada: ACM, 2014. 2672−2680

13 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of the 2016 International Conference on Learning Representations (ICLR). Caribe Hilton, San Juan, Puerto Rico, 2016.

14 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 1411.1784, 2014. 2672−2680

15 Denton E, Chintala S, Szlam A, Fergus R. Deep generative image models using a Laplacian pyramid of adversarial networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS). Montreal, Canada: ACM, 2015. 1486−1494

16 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv: 1701.07875, 2017.

17 Agarwal A, Triggs B. 3D human pose from silhouettes by relevance vector regression. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC, USA: IEEE, 2004. II-882−II-888

18 Mori G, Malik J. Estimating human body configurations using shape context matching. In: Proceedings of the 7th European Conference on Computer Vision (ECCV). Copenhagen, Denmark: Springer, 2002. 666−680

19 Taylor G W, Fergus R, Williams G, Spiro I, Bregler C. Pose-sensitive embedding by nonlinear NCA regression. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS). Vancouver, BC, Canada: ACM, 2010. 2280−2288

20 Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005, **61**(1): 55−79

21 Jain A, Tompson J, Andriluka M, Taylor G, Bregler C. Learning human pose estimation features with convolutional networks. In: Proceedings of the 2014 ICLR. Banff, Canada, 2014. 1−14

22 Pfister T, Charles J, Zisserman A. Flowing convnets for human pose estimation in videos. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1913−1921

23 Jing X Y, Zhu X K, Wu F, Hu R M, You X G, Wang Y H, et al. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. *IEEE Transactions on Image Processing*, 2017, **26**(3): 1363−1378

24 Zheng Z D, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 3774−3782

25 Qian X L, Fu Y W, Xiang T, Wang W X, Qiu J, Wu Y, et al. Pose-normalized image generation for person re-identification. arXiv: 1712.02225, 2018.

26 Deng W J, Zheng L, Ye Q X, Kang G L, Yang Y, Jiao J B. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 994−1003

27 Mathieu M, Couprie C, LeCun Y. Deep multi-scale video prediction beyond mean square error. In: Proceedings of the 4th International Conference on Learning Representations (ICLR). Caribe Hilton, San Juan, Argentina, 2016.

28 Wei S E, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 4724−4732

29 Ramakrishna V, Munoz D, Hebert M, Bagnell J A, Sheikh Y. Pose machines: articulated pose estimation via inference machines. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 33−47

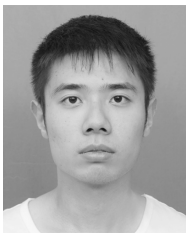
30 Cao Z, Simon T, Wei S E, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1302−1310

刘一敏 合肥工业大学计算机与信息学院硕士研究生. 主要研究方向为计算机视觉, 图像处理, 行人再识别.
E-mail: yiminliu@mail.hfut.edu.cn
(**LIU Yi-Min** Master student at the School of Computer and Information, Hefei University of Technology. His research interest covers computer vision, image processing, and person re-identification.)

蒋建国 合肥工业大学计算机与信息学院教授. 主要研究方向为数字图像分析和处理, 分布式智能系统和数字信号处理技术及应用.
E-mail: jgjiang@hfut.edu.cn
(**JIANG Jian-Guo** Professor at the School of Computer and Information, Hefei University of Technology. His research interest covers digital image analysis and processing, distributed intelligent systems, digital signal processing (DSP) technology, and applications.)

齐美彬 合肥工业大学计算机与信息学院教授. 主要研究方向为视频编码, 运动目标检测与跟踪和 DSP 技术. 本文通信作者. E-mail: qimeibin@163.com
(**QI Mei-Bin** Professor at the School of Computer and Information, Hefei University of Technology. His research interest covers video coding, moving target detection and tracking, and DSP technology. Corresponding author of this paper.)

刘 皓 腾讯优图实验室研究员. 2018 年获得合肥工业大学博士学位. 主要研究方向为计算机视觉, 行人再识别, 图像检索. E-mail: hfut.haoliu@gmail.com
(**LIU HAO** Researcher of Tencent YouTu Laboratory, He received his Ph. D. degree from Hefei University of Technology in 2018. His research interest covers computer vision, person re-identification, and image retrieval.)



周华捷 合肥工业大学计算机与信息学院硕士研究生. 主要研究方向为计算机视觉, 图像处理, 行人再识别.
E-mail: Zhou_hj@mail.hfut.edu.cn
(**ZHOU Hua-Jie** Master student at the School of Computer and Information, Hefei University of Technology. His research interest covers computer vision, image processing, and person re-identification.)