分类号：_____ 单位代码：_____

学　　号：_____

浙江大学

博士学位论文开题报告

中文论文题目：基于行人重识别的跨摄像头多目标跟踪方法研究

英文论文题目：**Study on multi-target multi-camera tracking based on person re-identification**

姓名：罗浩

导师：姜伟

专业：控制科学与工程

学号：11532034

学院：控制学院

报告日期　2017年11月

摘　　要

关键词： 行人重识别，跨摄像头多目标跟踪，深度学习，卷积神经网络

目　　次

1 研究意义与背景

2 国内外研究现状

本章节主要介绍本课题相关的研究现状，包括行人重识别(Person re-identification, person ReID)和跨摄像头多目标跟踪(Multi-target multi-camera tracking, MTMC tracking)两个部分。在本章节将会分别介绍这两个子课题相关的数据集和现有算法。其中行人重识别着重介绍近几年深度卷积神经网络相关的方法，而跨摄像头多目标跟踪将会着重介绍基于行人重识别的方法。

2.1 行人重识别

行人重识别也称行人再识别，是利用计算机视觉技术判断图像或者视频序列中是否存在特定行人的技术。广泛被认为是一个图像检索的子问题。给定一个监控行人图像，检索跨设备下的该行人图像。旨在弥补目前固定的摄像头的视觉局限，并可与行人检测/行人跟踪技术相结合，可广泛应用于智能视频监控、智能安保等领域。

而对于跨摄像头多目标跟踪问题，当一个行人目标在其中一个摄像头视野中消失后，要把该行人在其他摄像头中再次识别出来，这就是典型的行人重识别问题。也就是说，行人重识别技术是跨摄像头多目标跟踪的基础。因此，在本小节将会先介绍现有的行人重识别相关的数据集、准确度评估准则和一些现有的主流方法。

2.1.1 相关数据集

行人重识别相关的数据集总共有十几个，在早年深度学习还未出现的时候，那时的数据集图片数量还比较少。随着深度学习的诞生，行人重识别问题对数据量的要求大大增加，本小节将介绍几个适用深度学习的大规模行人出识别数据集。

- Market1501

  Market1501[1]是在清华大学校园中采集，图像来自6个不同的摄像头，其中有一个摄像头为低像素。同时该数据集提供训练集和测试集。训练集包含12,936张图像，测试

集包含19,732 张图像。图像由检测器自动检测并切割，包含一些检测误差（接近实际使用情况）。训练数据中一共有751人，测试集中有750 人。所以在训练集中，平均每类（每个人）有17.2张训练数据。

- MARS

  MARS (Motion Analysis and Re-identification Set)[2]数据集是Market1501的扩展。该数据集的图像由检测器自动切割，包含了行人图像的整个跟踪序列(tracklet)。MARS总共提供1,267个行人的20,478个图像序列，和Market1501一样来自同样的6个摄像头。和其他单帧图像数据集不一样的地方是，MARS是提供序列信息的大规模行人重识别数据集。

- CUHK03

  CUHK03[3] 在香港中文大学采集，图像来自2个不同的摄像头。该数据集提供机器自动检测和手动检测两个数据集。其中检测数据集包含一些检测误差，更接近实际情况。数据集总共包括1,467个行人的14,097张图片，平均每个人有9.6张训练数据。

- CUHK-SYSU

  CUHK-SYSU[4]是香港中文大学和中山大学一起收集的数据集。该数据集的特点是提供整个完整的图片，而不像其他大部分数据集一样只提供自动或者手动提取边框(bounding box)的行人图片。该数据集总共包括18,184张完整图片，内含8,432个行人的99,809张行人图片。其中训练集有11,206张完整图片，包含5,532个行人。测试集有6,978张完整图片，包含2,900个行人。

- DukeMTMC-reID

  DukeMTMC-reID[5]在杜克大学内采集，图像来自8个不同摄像头，行人图像的边框由人工标注完成。该数据集提供训练集和测试集。训练集包含16,522张图像，测试集包含17,661 张图像。训练数据中一共有702人，平均每个人有23.5 张训练数据。该数据集是目前最大的行人重识别数据集，并且提供了行人属性（性别/长短袖/是否背包等）的标注。

- VIPeR

  VIPeR[6]数据集是早期的一个小型行人重识别数据集，图像来自2个摄像头。该数据集总共包含632个行人的1,264，每个行人有两张不同摄像头拍摄的图片。数据集随

机分为相等的两部分，一部分作为训练集，一部分作为测试集。由于采集时间较早，该数据集的图像分辨率非常低，所以识别难度较大。

- PRID2011

  PRID2011[7]是2011年提出的一个数据集，图像来自于2个不同的摄像头。该数据集总共包含934个行人的24,541张行人图片，所以的检测框都是人工手动提取。图像大小的分辨率统一为128×64的分辨率。

以上是目前行人重识别研究中主要运用的数据集。由于行人重识别图片采自于不同摄像头，所以会出现光照、行人姿态、拍摄角度、遮挡、图像模糊等问题，造成同一行人的图片在不同摄像头中表现差异很大。如图2-1所示，上一排与下一排为同一个行人在两个不同摄像头拍摄的图片。可以看出，第一列存在遮挡现象，第二列至第四列存在拍摄角度、姿态等的巨大差异，第五列由于拍摄距离不同造成行人占图像比例大小差异很大，而最后一列是典型的摄像头分辨率不同而造成的图像差异。正式因为各种因素造成的图像差异，所以使得行人重识别很难通过手动提取特征就达到很好的识别效果，需要通过一定手段来学习到非常鲁棒的图像特征。

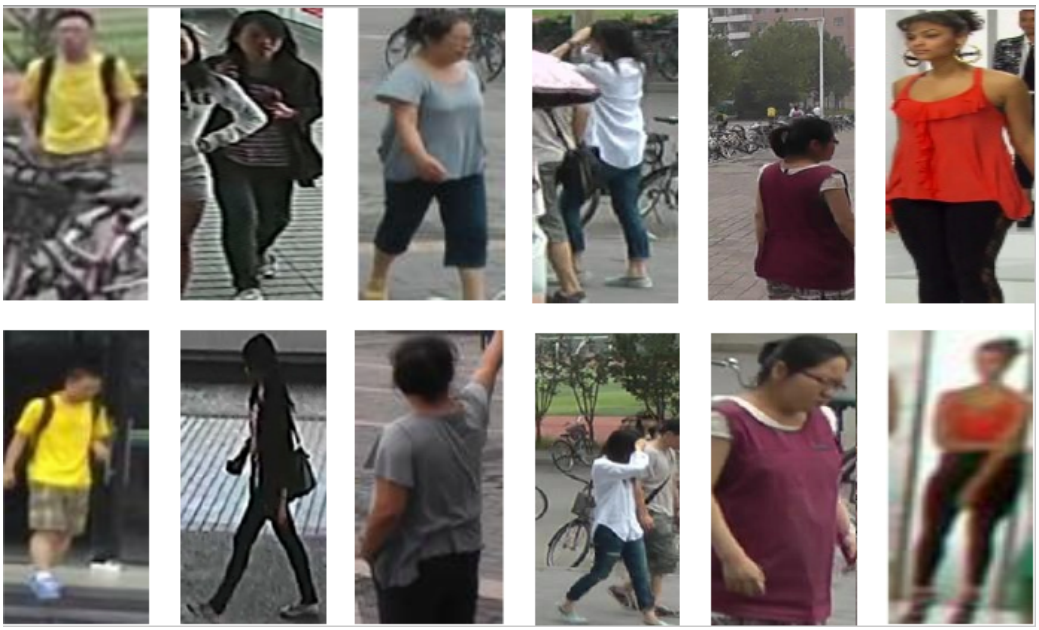

图 2-1 行人重识别数据集图片示例

2.1.2 准确度评估准则

为了评估行人重识别算法的优劣，需要统一一些评价准则。在学术论文中，通常大家默认选择累计匹配(Cumulative Match Characteristics, CMC)曲线和平均准确度(Mean Average

Precision, mAP)来作为评价准则。CMC和mAP是检索问题中常用的评价准则，在介绍它们之前，我们先介绍一些要用到常用术语。

- query：指测试集中的待检索库,包含图片的数目为N_q。
- gallery：指测试集中的搜索库。
- probe：指query中的某张待检索的图片，测试时需要将gallery中和probe为同一行人的图片全部检索出来。

(1) CMC曲线

CMC曲线主要用于计算rank-k的击中概率，在行人重识别、人脸识别领域使用较多。针对于query集中的一张带检索的probe图片，返回gallery的一系列排好序的结果，排序按照相似度排序。越靠前的结果表示和probe图片越相似，在行人重识别领域也等同于和probe是同一个人的概率越高。在测试阶段，需要排除gallery集中和probe处于同一摄像头的图片，防止其参与检索排序。我们设$index_{probe}$表示和gallery和probe为相同行人的最靠前的排序结果。最后rank-k准确度$A(rank\text{-}k)$可以表示为：

$$A(rank\text{-}k) = \frac{\sum_{probe \in query} f_{CMC}(index_{probe}, k)}{N_q} \tag{2-1}$$

其中：

$$f_{CMC}(index_{probe}, k) = \begin{cases} 0 & index_{probe} > k \\ 1 & index_{probe} \leq k \end{cases} \tag{2-2}$$

在实际使用中，为了减少计算量，通常我们比较关心rank-1,rank-5,rank-10,rank-20等准确度。

(2) mAP

mAP是另外一种重要的评价指标。CMC曲线通常只关心检索库中最靠前的正样本排序，而mAP由gallery中所有正样本的排序结果决定，所以通常能够更加鲁邦地反映模型的性能。计算mAP需要以下三步：

（1）Precision：对于query中的某一张probe图片，返回了gallery的一系列排序结果，考虑前n个查询结果，$P(n)$=前n个结果中与probe图片是相同行人的数目/n；

（2）Average Precision：对于query的第K个probe图片，记录排序结果中所有M个正样本排序结果的集合$\{i_1, i_2, ..., i_M\}$，计算它们的平均Precision，即$AP_K = \sum P(i)/M$，其中$i \in \{i_1, i_2, ..., i_M\}$；

（3）Mean Average Precision (mAP)：所有N_q张probe图片的Average Precision 的平均值，即$mAP = \sum_K AP_K/N$。

2.1.3 基于表征学习的方法

基于表征学习(Representation learning)的方法是一类非常常用的行人重识别方法[8–11]。这主要得益于深度学习，尤其是卷积神经网络(Convolutional neural network, CNN)[12]的快速发展。由于CNN可以自动从原始的图像数据中根据任务需求自动提取出表征特征(Representation)，所以有些研究者把行人重识别问题看做分类(Classification/Identification)问题或者验证(Verification)问题。分类问题是指利用行人的ID或者属性等作为训练标签来训练模型。验证问题是指输入一对（两张）行人图片，让网络来学习这两张图片是否属于同一个行人。

论文[8]利用Classification/Identification loss和verification loss来训练网络，其网络示意图如图2-2所示。网络输入为若干对行人图片，包括分类子网络(Classification Subnet)和验证子网络(Verification Subnet)。分类子网络对图片进行ID预测，根据预测的ID来计算分类误差损失。验证子网络融合两张图片的特征，判断这两张图片是否属于同一个行人，该子网络实质上等于一个二分类网络。经过足够数据的训练，再次输入一张测试图片，网络将自动提取出一个特征，这个特征用于行人重识别任务。

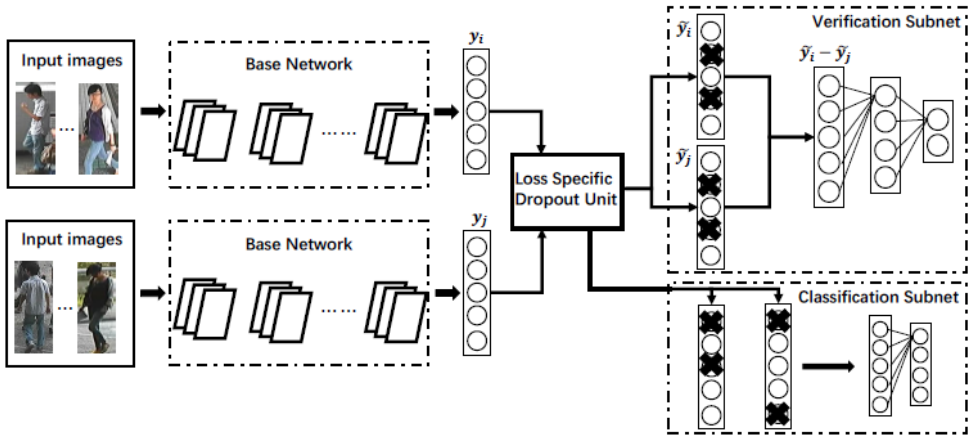

Figure 1. The proposed deep Re-ID network architecture.

图 2-2 结合分类损失和验证损失训练ReID网络示意图

论文[9–11]认为光靠行人的ID信息不足以学习出一个泛化能力足够强的模型。在这些工作中，它们额外标注了行人图片的属性特征，例如性别、头发、衣着等属性。通过引入行人属性标签，模型不但要准确地预测出行人ID，还要预测出各项正确的行人属性，这大大增加了模型的泛化能力，多数论文也显示这种方法是有效的。图2-3是其中一个示例，从

图中可以看出，网络输出的特征不仅用于预测行人的ID信息，还用于预测各项行人属性。通过结合ID损失和属性损失能够提高网络的泛化能力。

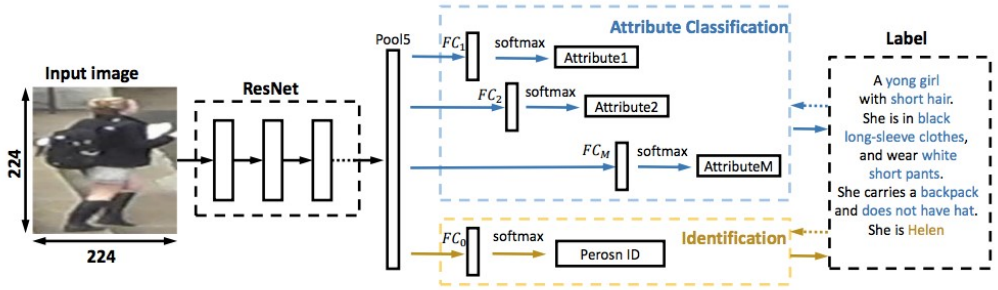

Figure 2. An overview of the APR network. During training, it predicts M attribute labels and an ID label. The weighted sum of the individual losses is back propagated. During testing, we extract the Pool5 (ResNet-50) or FC7 (CaffeNet) descriptors for retrieval.

图 2-3 结合行人ID标注和行人属性训练ReID网络示例

2.1.4 基于度量学习的方法

度量学习(Metric learning)是广泛用于图像检索利于的一种方法。不同于表征学习，度量学习旨在通过网络学习出两张图片的相似度。在行人重识别问题上，具体为同一行人的不同图片相似度大于不同行人的不同图片。最后网络的损失函数使得相同行人图片（正样本对）的距离尽可能小，不同行人图片（负样本对）的距离尽可能大。常用的度量学习损失方法有对比损失(Contrastive loss)[13]、三元组损失(Triplet loss)[14–16]、四元组损失(Quadruplet loss)[17]。首先，假如有两张输入图片I_1和I_2，通过网络的前馈我们可以得到它们归一化后的特征向量f_{I_1}和f_{I_2}。我们定义这两张图片特征向量的欧式距离为：

$$d_{I_1,I_2} = ||f_{I_1} - f_{I_2}||_2 \tag{2-3}$$

(1) 对比损失(Contrastive loss)

对比损失用于训练孪生网络(Siamese network)，其结构图如图2-4所示。孪生网络的输入为一对（两张）图片I_a和I_b，这两张图片可以为同一行人，也可以为不同行人。每一对训练图片都有一个标签y，其中$y=1$表示两张图片属于同一个行人（正样本对），反之$y=0$表示它们属于不同行人（负样本对）。之后，对比损失函数写作:

$$L_c = y d_{I_a,I_b}^2 + (1-y)(\alpha - d_{I_a,I_b})_+^2 \tag{2-4}$$

其中$(z)_+$表示$max(z,0)$，α是根据实际需求设计的阈值参数。为了最小化损失函数，当网络输入一对正样本对，$d(I_a,I_b)$会逐渐变小，即相同ID的行人图片会逐渐在特征空间形成聚类。反之，当网络输入一对负样本对时，$d(I_a,I_b)$会逐渐变大直到超过设定的α。

通过最小化L_c，最后可以使得正样本对之间的距离逐渐变下，负样本对之间的距离逐渐变大，从而满足行人重识别任务的需要。

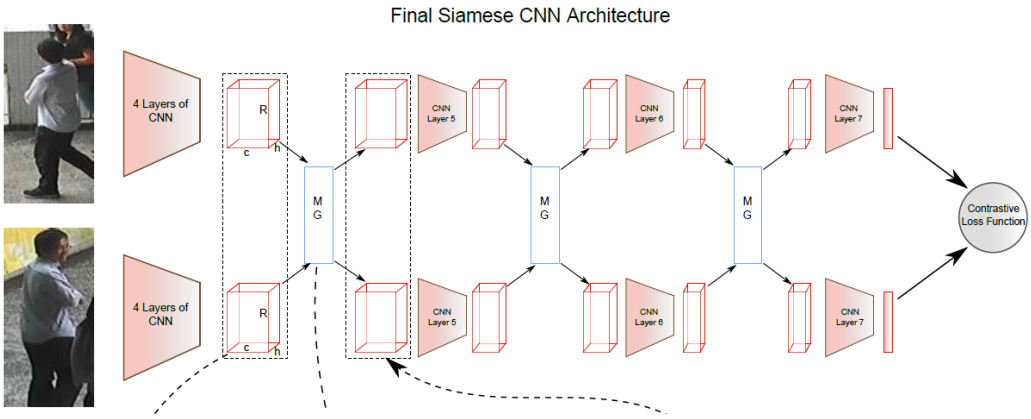

图 2-4 孪生网络结构示意图

(2) 三元组损失(Triplet loss)

三元组损失是一种被广泛应用的度量学习损失，之后的大量度量学习方法也是基于三元组损失演变而来。顾名思义，三元组损失需要三张输入图片。和对比损失不同，一个输入的三元组（Triplet）包括一对正样本对和一对负样本对。三张图片分别命名为固定图片(Anchor) a，正样本图片(Positive) p和负样本图片(Negative) n。图片a和图片p为一对正样本对，图片a和图片n为一对负样本对。则三元组损失表示为：

$$L_t = (d_{a,p} - d_{a,n} + \alpha)_+ \tag{2-5}$$

如图2-5所示，三元组可以拉近正样本对之间的距离，推开负样本对之间的距离，最后使得相同ID的行人图片在特征空间里形成聚类，达到行人重识别的目的。

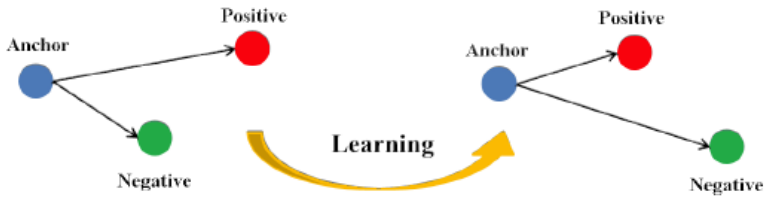

Figure 4. Triplet loss.

图 2-5 三元组损失[18]

论文[16]认为公式(2-5)只考虑正负样本对之间的相对距离，而并没有考虑正样本对之间的绝对距离，为此提出改进三元组损失(Improved triplet loss)：

$$L_i t = d_{a,p} + (d_{a,p} - d_{a,n} + \alpha)_+ \tag{2-6}$$

公式(2-6)添加$d_{a,p}$项，保证网络不仅能够在特征空间把正负样本推开，也能保证正样本对之间的距离很近。

(3) 四元组损失(Quadruplet loss)

四元组损失是三元组损失的另一个改进版本。顾名思义，四元组(Quadruplet)需要四张输入图片，和三元组不同的是多了一张负样本图片。即四张图片为固定图片(Anchor) a，正样本图片(Positive) p,负样本图片1(Negative1) $n1$ 和负样本图片2(Negative2) $n2$。其中$n1$和$n2$是两张不同行人ID的图片，其结构如图2-6所示。则，四元组损失表示为：

$$L_t = (d_{a,p} - d_{a,n1} + \alpha)_+ + (d_{a,p} - d_{n1,n2} + \beta)_+ \tag{2-7}$$

其中α和β是手动设置的正常数，通常设置β小于α，前一项称为强推动，后一项称为弱推动。相比于三元组损失只考虑正负样本间的相对距离，四元组添加的第二项不共享ID，所以考虑的是正负样本间的绝对距离。因此，四元组损失通常能让模型学习到更好的表征。

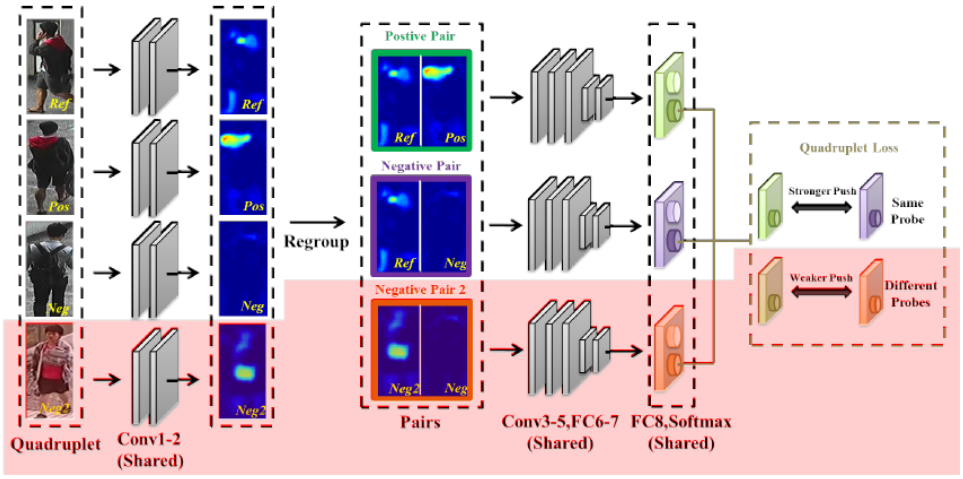

Figure 3. The framework of the proposed quadruplet deep network. The red shadow region indicates elements of the new constraint.

图 2-6 四元组损失网络结构图

(4) 难样本采样三元组损失(Triplet loss with hard sample mining)

难样采样三元组损失（本文之后用TriHard损失表示）是三元组损失的改进版。传统的三元组随机从训练数据中抽样三张图片，这样的做法虽然比较简单，但是抽样出来的大部分都是简单易区分的样本对。如果大量训练的样本对都是简单的样本对，那么这是不利于网络学习到更好的表征。大量论文发现用更难的样本去训练网络能够提高网络的泛化能力，而采样难样本对的方法很多。论文[19]提出了一种基于训练批量(Batch)的在线难样本采样方法——TriHard损失。

TriHard损失的核心思想是：对于每一个训练batch，随机挑选P个ID的行人，每个行人随机挑选K张不同的图片，即一个batch含有$P \times K$张图片。之后对于batch中的每一张图片a，我们可以挑选一个最难的正样本和一个最难的负样本和a组成一个三元组。

首先我们定义和a为相同ID的图片集为A，剩下不同ID的图片图片集为B，则TriHard损失表示为：

$$L_{th} = \frac{1}{P \times K} \sum_{a \in batch} (\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha)_+ \tag{2-8}$$

其中α是人为设定的阈值参数。TriHard损失会计算a和batch中的每一张图片在特征空间的欧式距离，然后选出与a距离最远（最不像）的正样本p和距离最近（最像）的负样本n来计算三元组损失。通常TriHard损失效果比传统的三元组损失要好。

2.1.5 基于局部特征的方法

从网络的训练损失函数上进行分类可以分成表征学习和度量学习，相关方法前文已经介绍。另一个角度，从抽取图像特征进行分类，行人重识别的方法可以分为基于全局特征(Global feature) 和基于局部特征(Local feature)的方法。全局特征是指让网络对整幅图像提取一个特征，这个特征不考虑一些局部信息。而局部特征是指让手动或者自动地让网络去关注关键的局部区域，然后提取这些区域的局部特征。常用的提取局部特征的思路主要有图像切块、利用骨架关键点定位以及姿态矫正等等。

图片切块是一种很常见的提取局部特征方式[20,21]。如图2-7所示，图片被垂直等分为若干份，因为垂直切割更符合我们对人体识别的直观感受，所以行人重识别领域很少用到水平切割。之后，被分割好的若干块图像块按照顺序送到一个长短时记忆网络(Long short term memory network, LSTM)，最后的特征融合了所有图像块的局部特征。但是这种缺点在于对图像对齐的要求比较高，如果两幅图像没有上下对齐，那么很可能出现头和上身对比的现象，反而使得模型判断错误。

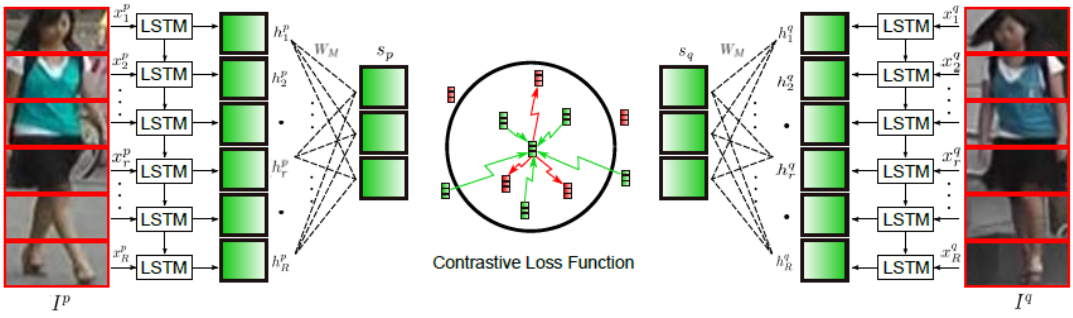

图 2-7 利用图片切块提取局部特征示例

为了解决图像不对齐情况下手动图像切片失效的问题，一些论文利用一些先验知识先将行人进行对齐，这些先验知识主要是预训练的人体姿态(Pose)和骨架关键点(Skeleton)模型。论文[22]先用姿态估计的模型估计出行人的关键点，然后用仿射变换使得相同的关键点对齐。如图2-8所示，一个行人通常被分为14个关键点，这14个关键点把人体结果分为若干个区域。为了提取不同尺度上的局部特征，作者设定了三个不同的PoseBox组合。之后这三个PoseBox矫正后的图片和原始为矫正的图片一起送到网络里去提取特征，这个特征包含了全局信息和局部信息。特别提出，如果这个仿射变换可以在进入网络之前的预处理中进行，也可以在输入到网络后进行。如果是后者的话需要需要对仿射变换做一个改进，因为传统的放射变化是不可导的。为了使得网络可以训练，需要引入可导的近似放射变化，在本文中不赘述相关知识。

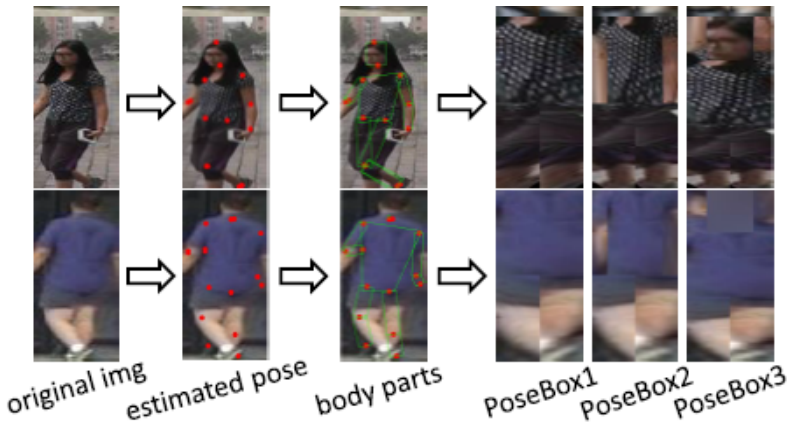

图 2-8 姿态对齐示意图

CVPR2017的工作Spindle Net[23]也利用了14个人体关键点来提取局部特征。和论文[22]不同的是，Spindle Net并没有用仿射变换来对齐局部图像区域，而是直接利用这些关键点来抠出感兴趣区域(Region of interest, ROI)。Spindle Net网络如图2-9所示，首先通过骨架关键点提取的网络提取14个人体关键点，之后利用这些关键点提取7个人体结构ROI。网络中所有提取特征的CNN（橙色表示）参数都是共享的，这个CNN分成了线性的三个子网络FEN-C1、FEN-C2、FEN-C3。对于输入的一张行人图片，有一个预训练好的骨架关键点提取CNN（蓝色表示）来获得14个人体关键点，从而得到7个ROI区域，其中包括三个大区域（头、上身、下身）和四个四肢小区域。这7个ROI区域和原始图片进入同一个CNN网络提取特征。原始图片经过完整的CNN得到一个全局特征。三个大区域经过FEN-C2和FEN-C3子网络得到三个局部特征。四个四肢区域经过FEN-C3子网络得到四个局部特征。之后这8个特征按照图示的方式在不同的尺度进行联结，最终得到一个融合全局特征和多个尺度局部特征的行人重识别特征。

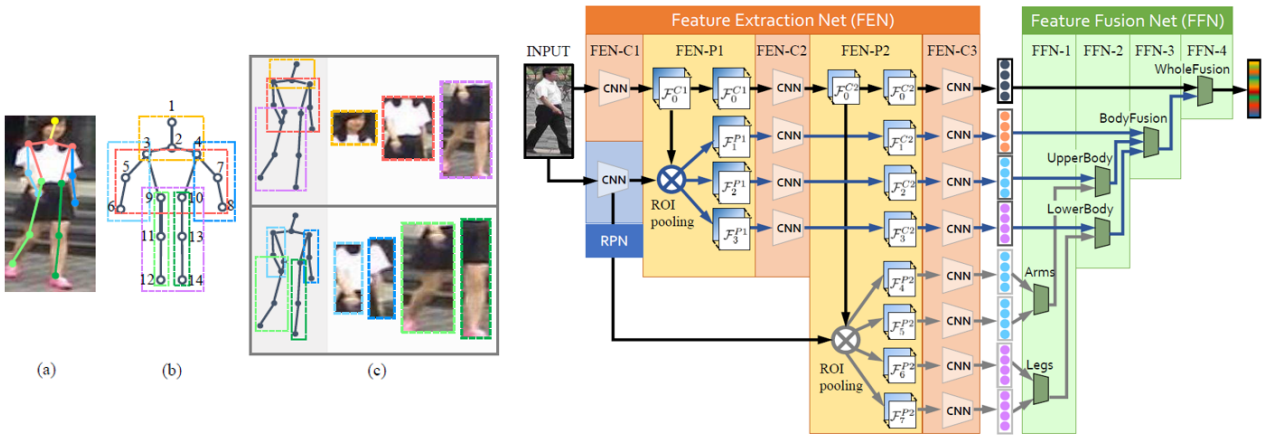

图 2-9　Spindle Net结构示意图

论文[24]提出了一种全局-局部对齐特征描述子(Global-Local-Alignment Descriptor, GLAD)，来解决行人姿态变化的问题。与Spindle Net类似，GLAD利用提取的人体关键点把图片分为头部、上身和下身三个部分。之后将整图和三个局部图片一起输入到一个参数共享CNN网络中，最后提取的特征融合了全局和局部的特征。为了适应不同分辨率大小的图片输入，网络利用全局平均池化(Global average pooling, GAP)来提取各自的特征。和Spindle Net略微不同的是四个输入图片各自计算对应的损失，而不是融合为一个特征计算一个总的损失。

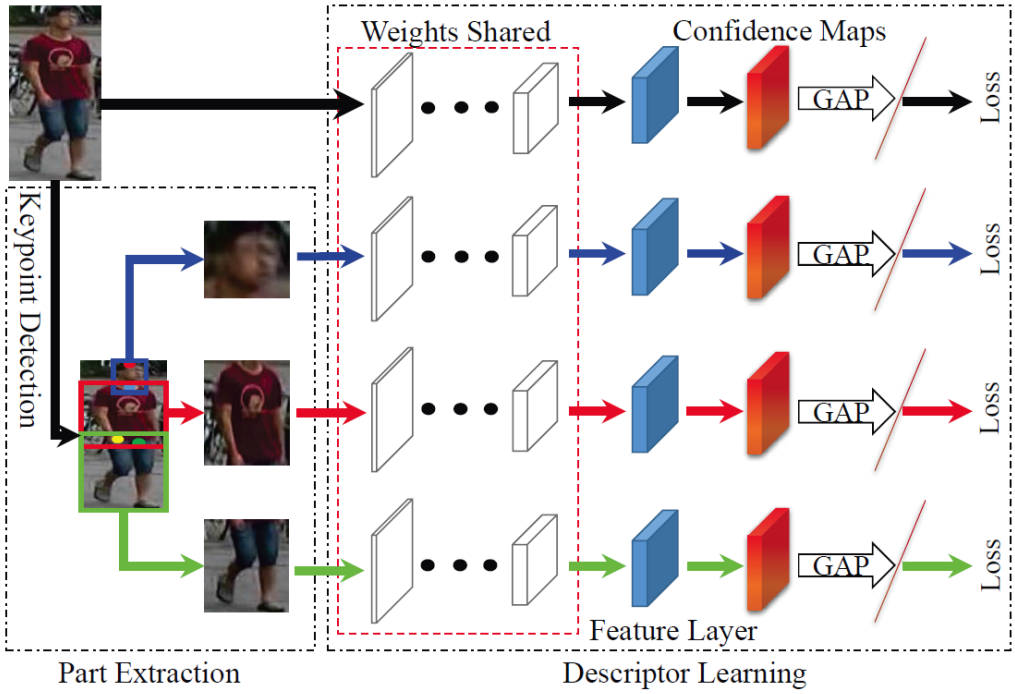

图 2-10　GLAD结构示意图

2.1.6 基于视频序列的方法

以上介绍的方法都是基于单帧图像的方法，通常单帧图像的信息是有限的，因此有很多工作集中在利用视频序列来进行行人重识别方法的研究[25–31]。基于视频序列的方法最主要的不同点就是这类方法不仅考虑了图像的内容信息，还考虑了帧与帧之间的运动信息等。

基于单帧图像的方法主要思想是利用CNN来提取图像的空间特征，而基于视频序列的方法主要思想是利用CNN 来提取空间特征的同时利用递归循环网络(Recurrent neural networks, RNN)来提取时序特征。图2-11是非常典型的思路，网络输入为图像序列。每张图像都经过一个共享的CNN提取出图像空间内容特征，之后这些特征向量被输入到一个RNN网络去提取最终的特征。最终的特征融合了单帧图像的内容特征和帧与帧之间的运动特征。而这个特征用于代替前面单帧方法的图像特征来训练网络。

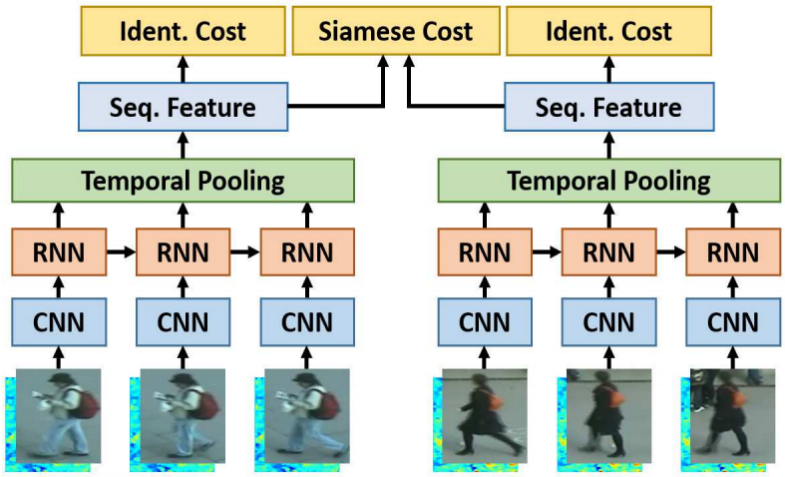

图 2-11 基于视频序列的行人重识别网络结构示意图

视频序列类的代表方法之一是累计运动背景网络(Accumulative motion context network, AMOC)[31]。AMOC输入的包括原始的图像序列和提取的光流序列。通常提取光流信息需要用到传统的光流提取算法，但是这些算法计算耗时，并且无法与深度学习网络兼容。为了能够得到一个自动提取光流的网络，作者首先训练了一个运动信息网络(Motion network, Moti Nets)。这个运动网络输入为原始的图像序列，标签为传统方法提取的光流序列。如图2-12所示，原始的图像序列显示在第一排，提取的光流序列显示在第二排。网络有三个光流预测的输出，分别为Pred1，Pred2，Pred3，这三个输出能够预测三个不同尺度的光流图。最后网络融合了三个尺度上的光流预测输出来得到最终光流图，预测的光流序列在第三排显示。通过最小化预测光流图和提取光流图的误差，网络能够提取出较准确的运动特征。

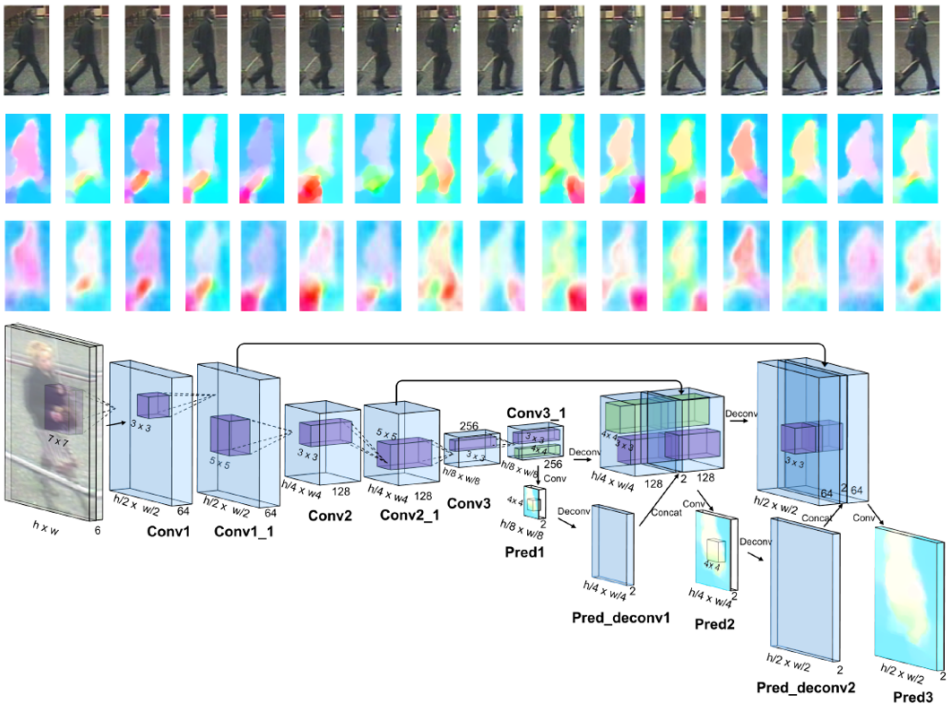

图 2-12　运动网络结构示意图

AMOC的核心思想在于网络除了要提取序列图像的特征，还要提取运动光流的运动特征，其网络结构图如图2-13所示。AMOC拥有空间信息网络(Spatial network, Spat Nets)和运动信息网络两个子网络。图像序列的每一帧图像都被输入到Spat Nets来提取图像的全局内容特征。而相邻的两帧将会送到Moti Nets来提取光流图特征。之后空间特征和光流特征融合后输入到一个RNN来提取时序特征。通过AMOC网络，每个图像序列都能被提取出一个融合了内容信息、运动信息的特征。网络采用了分类损失和对比损失来训练模型。融合了运动信息的序列图像特征能够提高行人重识别的准确度。

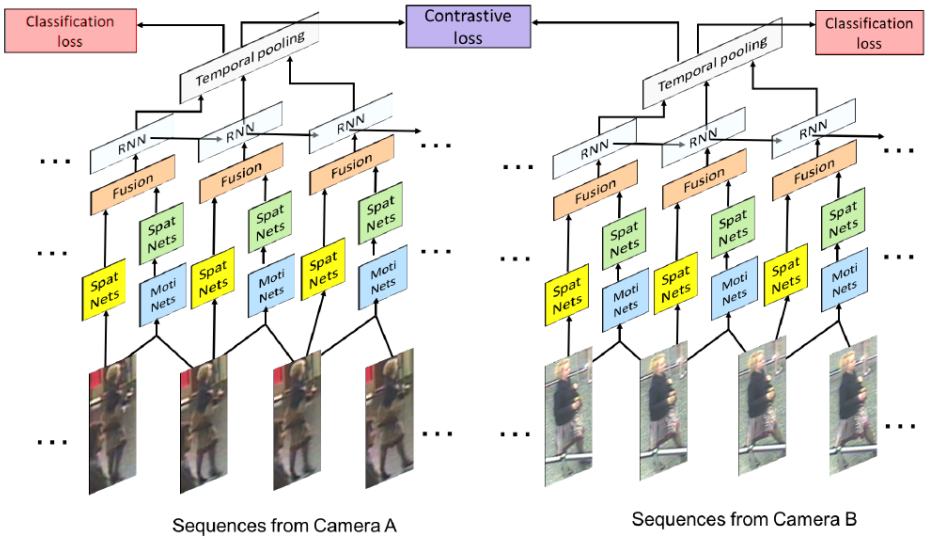

图 2-13　AMOC结构示意图

2.2 跨摄像头多目标跟踪

2.2.1 相关数据集

3 研究内容与技术路线

4 研究计划与现有成果

参考文献

[1] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, Qi Tian. Scalable person re-identification: A benchmark[C]//Computer Vision, IEEE International Conference. 2015.

[2] Springer. MARS: A Video Benchmark for Large-Scale Person Re-identification[J], 2016, 2016.

[3] Wei Li, Rui Zhao, Tong Xiao, Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification[J]. 2014:152–159.

[4] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, Xiaogang Wang. End-to-end deep learning for person search[J]. arXiv preprint arXiv:1604.01850, 2016.

[5] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking. 2016.

[6] Doug Gray, Shane Brennan, Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking[J]. 2007.

[7] Martin Hirzer, Csaba Beleznai, Peter M. Roth, Horst Bischof. Person re-identification by descriptive and discriminative classification[C]//Scandinavian Conference on Image Analysis. 2011:91–102.

[8] Mengyue Geng, Yaowei Wang, Tao Xiang, Yonghong Tian. Deep transfer learning for person re-identification[J]. arXiv preprint arXiv:1611.05244, 2016.

[9] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Yi Yang. Improving person re-identification by attribute and identity learning[J]. arXiv preprint arXiv:1703.07220, 2017.

[10] Liang Zheng, Yi Yang, Alexander G Hauptmann. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.

[11] Tetsu Matsukawa, Einoshin Suzuki. Person re-identification using cnn features learned from combination of attributes[C]//Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016:2428–2433.

[12] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. 2012:1097–1105.

[13] Rahul Rama Varior, Mrinal Haloi, Gang Wang. Gated siamese convolutional neural network architecture for human re-identification[C]//European Conference on Computer Vision. Springer, 2016:791–808.

[14] Florian Schroff, Dmitry Kalenichenko, James Philbin. Facenet: A unified embedding for face recogni-

tion and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:815–823.

[15] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, Shuicheng Yan. End-to-end comparative attention networks for person re-identification[J]. IEEE Transactions on Image Processing, 2017.

[16] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1335–1344.

[17] Weihua Chen, Xiaotang Chen, Jianguo Zhang, Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification[J]. arXiv preprint arXiv:1704.01719, 2017.

[18] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles[C]//Computer Vision and Pattern Recognition. 2016:2167–2175.

[19] Alexander Hermans, Lucas Beyer, Bastian Leibe. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[20] Qiqi Xiao, Kelei Cao, Haonan Chen, Fangyue Peng, Chi Zhang. Cross domain knowledge transfer for person re-identification[J]. arXiv preprint arXiv:1611.06026, 2016.

[21] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, Gang Wang. A siamese long short-term memory architecture for human re-identification[C]//European Conference on Computer Vision. Springer, 2016:135–153.

[22] Liang Zheng, Yujia Huang, Huchuan Lu, Yi Yang. Pose invariant embedding for deep person re-identification[J]. arXiv preprint arXiv:1701.07732, 2017.

[23] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]. CVPR, 2017.

[24] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval[J]. arXiv preprint arXiv:1709.04329, 2017.

[25] Taiqing Wang, Shaogang Gong, Xiatian Zhu, Shengjin Wang. Person re-identification by discriminative selection in video ranking[J]. IEEE transactions on pattern analysis and machine intelligence, 2016. 38(12):2501–2514.

[26] Dongyu Zhang, Wenxi Wu, Hui Cheng, Ruimao Zhang, Zhenjiang Dong, Zhaoquan Cai. Image-to-video person re-identification with temporally memorized similarity learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[27] Jinjie You, Ancong Wu, Xiang Li, Wei-Shi Zheng. Top-push video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1345–1353.

[28] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, Yisheng Zhong. Person re-identification by unsupervised video matching[J]. Pattern Recognition, 2017. 65:197–210.

[29] Niall McLaughlin, Jesus Martinez del Rincon, Paul Miller. Recurrent convolutional network for video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1325–1334.

[30] Rui Zhao, Wanli Oyang, Xiaogang Wang. Person re-identification by saliency learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2017. 39(2):356–370.

[31] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, Jiashi Feng. Video-based person re-identification with accumulative motion context[J]. arXiv preprint arXiv:1701.00193, 2017.