关于行人重识别技术的研究

章雯芳
21821146
3140103574@zju.edu.cn

摘要（Abstract）

本文主要对基于视频的行人重识别方法进行研究，相比于传统的单帧识别，基于视频的行人重识别方法可以在时空两个维度串联信息。但是视频识别也存在一些问题，比如对象的部分区域可能在某些帧内会被遮挡，因此造成干扰。对此Li[1]等提出了一种多项正则化模型来解决这个问题。本文主要对该论文提出的方法进行深入学习，首先对该论文提出的模型结构进行细致的描述，之后再将其与近期新提出的一些方法进行对比，最后，针对该论文的一些问题进行讨论。

1. 引言

行人重识别(Person Re-identification, 简称ReID)是通过分析不同镜头捕捉到的不重叠的画面，对同一个行人的行为进行匹配的技术。近来，该技术由于广泛的应用价值，在学界中引发了广泛的研究。目前主要的挑战在于由于镜头视野的多样性、行人的姿势、灯光以及遮挡等引发的问题。

已有的方法主要是将每帧进行独立编码，之后利用池化层来将不同帧之间的信息串联起来。这种方法无法解决有遮挡的样本的ReID问题。并且由于行人的动作是不断在改变的，池化也无法解决不同帧之间存在的空间不对齐问题。

文献[1]提出了SpaAtn模型。该模型的目的是在基于图片的ReID任务基础上进行泛化，优化的重点是能够挖掘出具有优秀表现能力的潜在特征。并且为了解决上述问题，该研究使用了时空注意力模型。其中空间注意力模型解决不同帧之间的对齐问题。其中，为了避免不同的空间模型学习同一个位置的特征，还设计了多样化的正则化项来避免这个问题。时间注意力模型则用来将不同帧的同一块区域串联起来。

2. 模型概述

模型架构如图 1所示。对于给定的一个视频序列，首先使用严格随机采样方法，选取视频中的部分帧作为训练样本。之后使用多个空间注意力模型，对这些样本进行学习，其中每个空间注意力模型都将学习躯干的不同部位对应的区域。之后，再使用时间注意力模型，分别对上一步学到的部位的区域做时序上的串联。最后，将揉合了时序信息的不同部位的特征连接起来，输入一个全连接层。这里使用的损失函数是OIM。

2.1. 严格随机采样

不同于单帧ReID问题，基于视频的ReID的学习样本是比较长的序列。然而相对来说，由于视频中连续的相邻帧都是高度相关的，在一个短区间范围内的帧并不会有明显的变化。对于比较长的区间，也就是距离比较远的帧，随着时间的推移，其中行人的动作可能产生比较大的变化。ReID问题主要就是要解决在长区间内，行人动作的变化情况。因此，可以采用严格随机采样的方法，这样即可以利用整个视频信息，又可以避免段区间内帧信息的大量冗余。

对于给定的一个视频，严格随机采样方法首先将其按照时间平均分为N块（对应上面的短区间），对于每个块，随机从中采样 1 帧。最后这个视频可以表达为由N张图像构成的集合。

2.2. 空间注意力

模型采用多个空间注意力模型来自动识别对ReID有用的图像区域（躯干和身上的其他装饰物）。相比于直接对图片进行网格化划分的刚性方法，采用多个空间

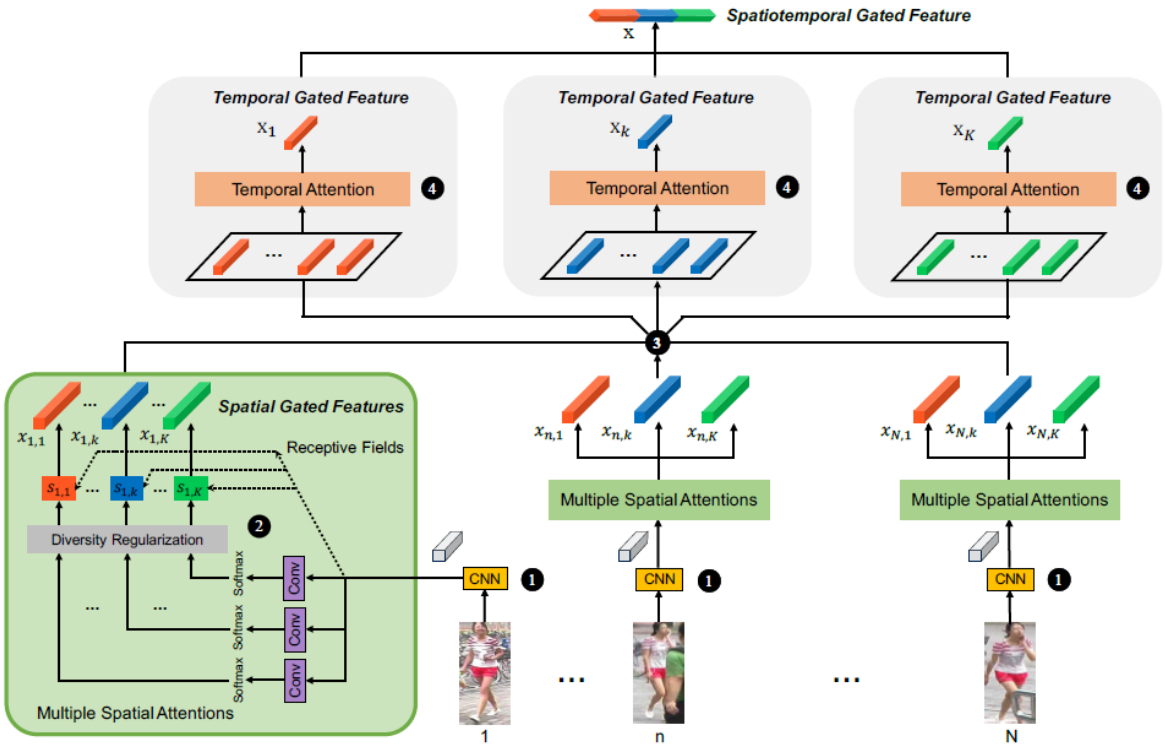

图 1 时空注意力模型结构

注意力模型能够自动识别每个图像中多个不相交，并且在训练视频中持续出现的特征区域。通过对这些区域的自动识别和定位，该方法能够解决由于动作、距离、遮挡引起的问题。并且，在空间注意力模型中，并不仅仅只是聚焦于躯干的某些部分，同样会对书包、帽子等配饰聚焦。直接从整个图像中生成的特征表示很容易错过细粒度的视觉线索(图 1)。该模型是第一个将空间注意力模型应用于视频中持续出现的图像帧中的区域识别的方法。

如图 1 所示，该方法首先使用一个ResNet-50 CNN架构作为基础模型，从图像中抽取特征，每个图像能够得到 32 个特征。之后采用k个空间注意力模型并行的对这些特征进行学习，从而得到显著特征（不同身体部分、装饰物）对应的区域信息。

为了确保不同的空间注意力模型学习的是不同的显著特征，该方法还引入了一个正则化项Q，Q可以看作是对每张图片之间显著区域的冗余的描述尺度。使用Q作为正则项，可以实现空间注意力模型关注区域的多样化，因此Q被成为多样化的正则项。

2.3. 时间注意力

如图 1 所示，通过空间注意力模型的学习，得到了同一个显著特征在不同帧中的区域信息，时间注意力模型即对这些在时间上呈现序列关系的特征进行表达学习。最后得到同一个显著区域在不同时间点行为的融合表达。

2.4. 模型输出

利用时间注意力模型，得到了k个显著特征贯穿整个视频的特征向量。在获得输出之前，我们对这k个向量进行简单的拼接，而后通过一个全连接层，将这个拼接向量映射到输出空间。

3. 实验

文通过在三个数据集——PRID2011, iLIDS-VID以及MARS上进行实验，其中MARS中的样本和实体数量比较多，而PRID2011 和iLIDS-VID都是较小的数据集。

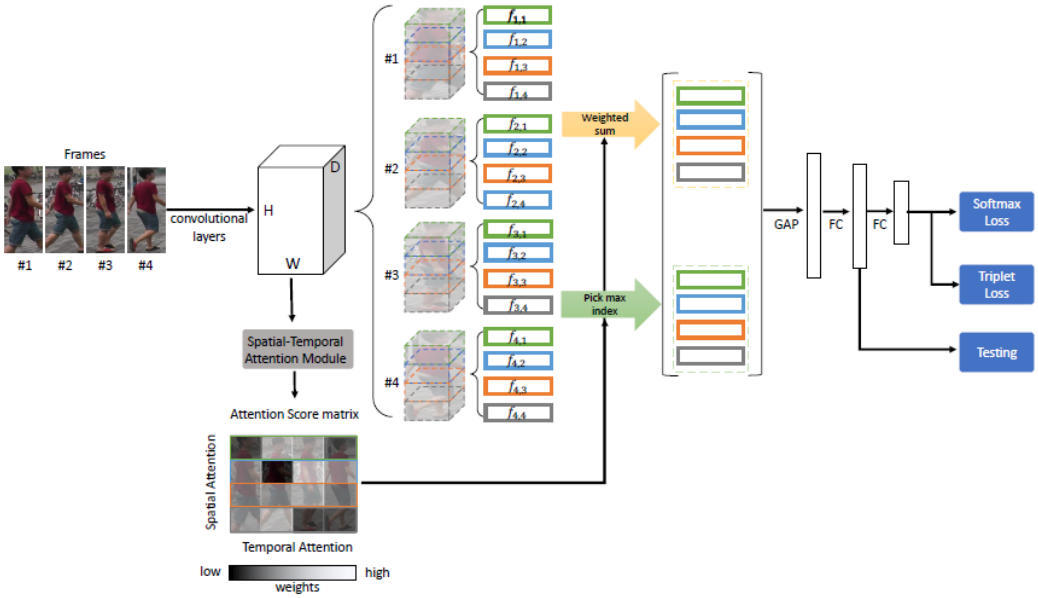

图 2 STA模型结构

通过实验，证明了该方法优于现有的其他方法。对本模型进行调整后，对比基准方法可以达到超过 10%的提升。

同时文章还对空间注意力模型的个数k进行了探究。随着k的增大，意味着模型可以注意到更多细节。但是通过实验可以得到，模型性能并非随着k的增大而随之提高，在k=6 时，模型达到了最优性能。背后的原因可能时由于当k增大时，正则项Q允许的显著区域间冗余越来越小，影响了模型性能的发挥。

4. 近期相似工作

Fu等[2]提出了STA方法，该方法针对同样的课题，使用类似的基于时空注意力模型，来解决视频ReID中存在的遮挡以及序列内各帧位置的不对齐问题。同样，STA抛弃以往工作中使用池化来进行帧间聚合的做法，而是选择使用空间注意力模型提供更加鲁棒的特征表示方法。STA的模型结构如图 2所示。不同于SpaAtn采用严格随机采样从视频中采样部分图像，STA直接在序列中随机进行采样。之后，将采样得到的图片输入一个ResNet-50 CNN网络中，提取每张图片对应的特征表示。在之后的空间注意力学习中，该模型使用了与SpaAtn截然不同的方法。SpaAtn为避免刚性划分带来的不灵活行，使用多样化的正则项Q使空间注意力模型自动学习不相交的显著区域。而STA将每张图像简单进行纵向的刚性划分，将其划分为 4 个区域。之后应用空间注意力模型，学习得到一个二维的注意力权重矩阵。之后通过不同的方法取得两种序列特征表示，分别是取每帧中权重最大的区域，以及对每帧的每个区域做加权和。最后，将两种序列特征表示输入全连接网络，通过softmax等方法，就学习得到了一个行人行为识别的模型。

5. 模型对比评价

将本文主要讨论的文献[1]提出的SpaAtn与STA进行对比，可以发现二者主要存在以下区别。

首先，在帧采样阶段。SpaAtn采取的采样方式更为严谨，可以较为全面的采样到视频序列中各个阶段。然而也因此，其输入的只能是定长的。而STA采用随机采样 4 张的方法，可以接受变长输入，但另一方面由于采样的随机性，可能发生采样不均衡的情况，从而遗漏行人的某些行为。

其次，在空间特征学习阶段。SpaAtn同时使用k个注意力模型对序列进行学习，自动识别追踪各个显著区域在不同图像中的位置，解决了不对齐问题。而STA使用简单的刚性划分方法，再利用空间注意力模型获得各个帧内部区域的注意力权重分布，以更有效率的方法解

决了空间不对齐问题。并且STA提出，SpaAtn使用k个不同的模型进行空间注意力学习的过程，可能忽略了区域之间的空间信息，例如人的身体的各个部位之间的位置关系。我的观点是，SpaAtn已经使用了大量的算力来自动追踪某个区域的变化位置，即使再添加相对位置关系的学习模块，能否对模型性能产生正向促进是存疑的，当然这需要实验来证明。

最后，在时序特征融合上。SpaAtn将每个显著区域单独进行学习，之后通过简单的连接将空间、时间特征融为一体。而STA直接同时使用时空模型，学习得到二维的注意力权重，在模型初始阶段即完成了时空融合。STA认为，分别使用空间和时间注意力模型进行学习，由于模型的不同，可能导致误差的积累。

最后，总体来说，SpaAtn相对STA来说具有更加复杂的模型结构，以及需要学习更多的参数。通过STA的实验，证明了其在大数据集上相比多项正则化时空注意力模型具有更好的性能。

6. 讨论

虽然文献[1]首次提出了将时空注意力模型应用与基于视频的ReID中，并且设计了非常精彩的模型框架，然而在模型细节方面还是有一些比较模糊的部分。在最后的实验结果中，提到了一个经过调整之后的框架，可以达到很好的实验效果，然而对于调整框架的哪些部分却缺少详细的说明，如果调整涉及到了基础的CNN部分，则也许在基准模型上的效果也可能能够得到相应的提升，因此该实验结果在说服力上有所欠缺。

其次，该文章公布在GitHub上的代码并不完整，他人难以复现结果。

References

[1] Li S, Bak S, Carr P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 369-378.

[2] Fu Y, Wang X, Wei Y, et al. STA: Spatial-Temporal Attention for Large-Scale Video-based Person Re-Identification[C]. AAAI, 2019.