GaitSet Regarding Gait as a Set for Cross-View Gait Recognition

摘要

现有的步态识别方法要么利用难以保留时间信息的步态模板，要么利用步态序列，后者必须保留不必要的顺序约束，从而失去了步态识别的灵活性步态研究中，时间序列信息是比较难描述的，

正文

介绍

文章提出了GaitSet这个网络，基于集合的思想，该网络不受帧排序的影响。

原本的步态模板方法简单易行，但容易丢失时间和细粒度的空间信息；近年很多直接从原始的步态轮廓序列中提取特征。但是，这些方法易受外部因素的影响。 此外，比起使用单个模板（如步态能量图像（GEI））的网络，像3D-CNN这样的用于提取顺序信息的深度神经网络更难训练。

文章使用GaitSet方法将一个集合的人体轮廓图送入网络，首先，使用CNN分别从每个轮廓提取帧级特征。其次，用称为集池化（Set Pooling）的操作用于将帧级特征聚合为单个集级特征。由于此操作是应用于高级特征图而不是原始轮廓，因此与步态模板相比，它可以更好地保留空间和时间信息。

GaitSet

一个N人数据集，每个人为y_i，$i \in 1,2,...,N$，某个人的剪影序列只与id号有关，记为P_i。一个人的一个或多个序列中的所有剪影都可以视为一组n个剪影$X_i = \{x_i^j | j = 1,2,...n\}$，$x_i^j \in P_i$，步态识别以以下公式表示：

$$f_i = H(G(F(X_i)))$$

其中F为用于提取帧级别的特征的卷积网络；G是置换不变函数，用集合池化(Set Pooling)实现，用于将一组帧级特征映射到集合级特征；H于从集合水平特征中学习P_i的判别式，用水平金字塔(Horizontal Pyramid Mappint, HMP)来实现；输入的X_i是一个4维张量(set dimension，image channel dimension，image hight dimension，image width dimension)

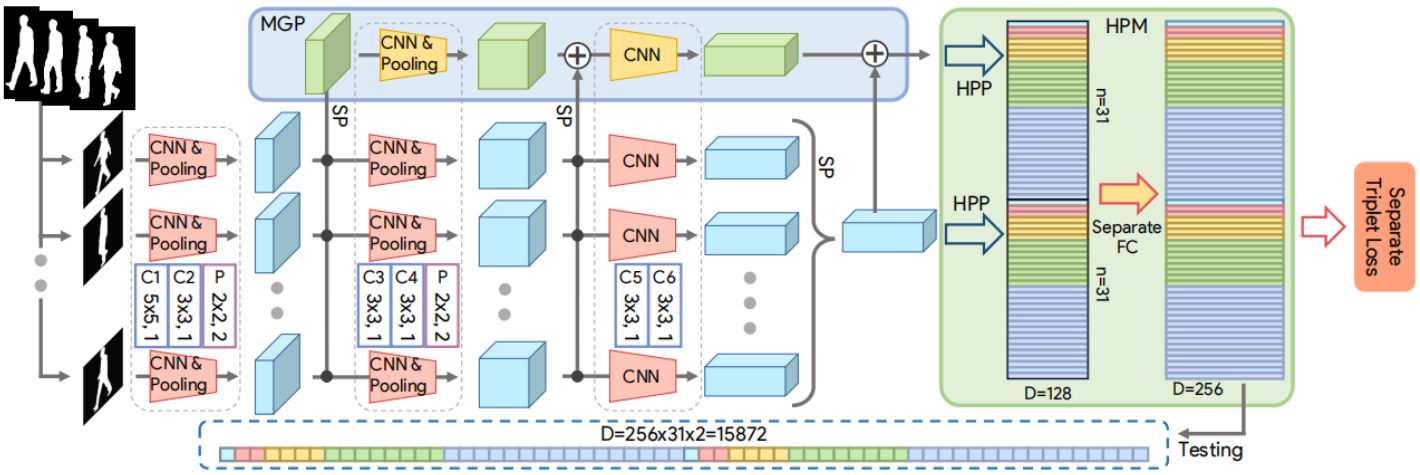

Figure 2: The framework of GaitSet. 'SP' represents Set Pooling. Trapezoids represent convolution and pooling blocks and those in the same column have the same configurations which are shown by rectangles with capital letters. Note that although blocks in MGP have same configurations with those in the main pipeline, parameters are only shared across blocks in the main pipeline but not with those in MGP. HPP represents horizontal pyramid pooling (Fu et al. 2018).

Set Pooling

集合池化整合一个集合的步态信息，$z = G(V)$，其中z表示集合级别特征，而$V = \{v^j | j = 1, 2, ..., n\}$表示帧级别特征。这个操作有两个约束，(1)是将集合作为输入，它应该是一个排列不变函数，其表达为：$G(\{v^j | j = 1, 2, ..., n\}) = G(\{v^{\pi(j)} | j = 1, 2, ..., n\})$，其中$\pi$为一个排列；(2)由于在现实生活中，一个人的步态剪影的数量可以是任意的，因此函数G应该可以处理具有任意基数的集合。

使用attention机制来利用全局信息为每个帧级特征图学习逐个元素的attention图，以对其进行细化。

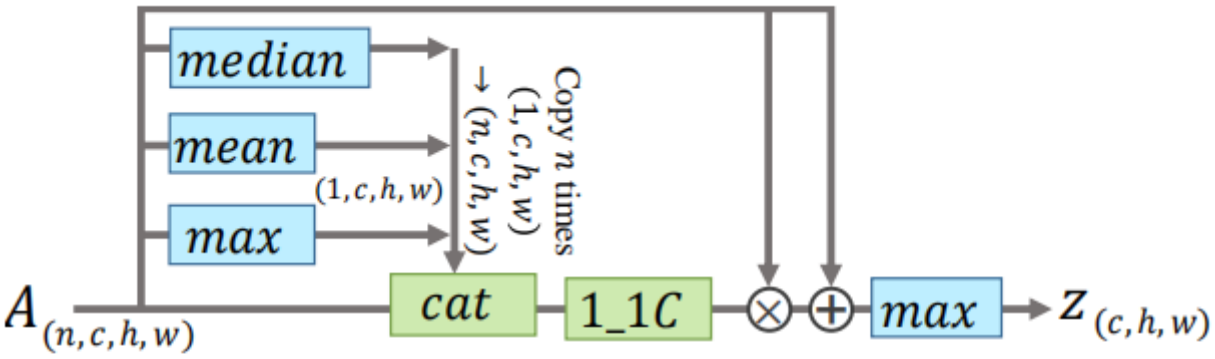

Figure 3: The structure of Set Pooling (SP) using attention. 1_1C and cat represents 1×1 convolutional layer and concatenate respectively. The multiplication and the addition are both pointwise.

将特征图分割成条带通常用于人员重新识别任务，图像被裁剪并根据行人的大小调整为统一大小，而区别部分随图像的不同而不同。文章将水平金字塔池化(Horizontal Pyramid Pooling, HPP)最后池化后面的$1*1$卷积层用全连接层替代，对每个合并要素使用独立的完全连接层（FC），以将其映射到区分性空间，这个方法被称为水平金字塔映射(Horizontal Pyramid Mapping, HPM)

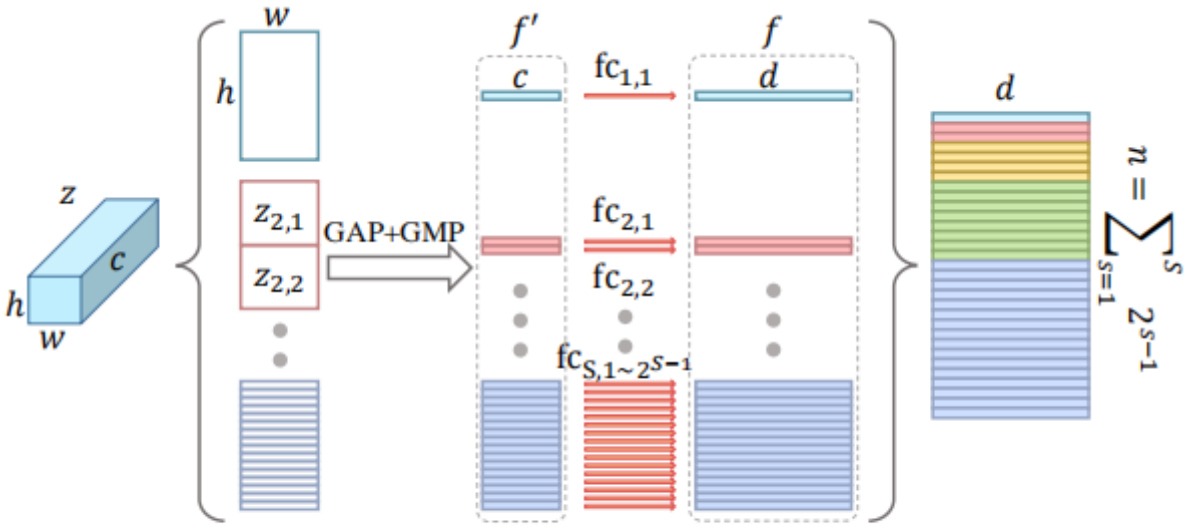

Figure 4: The structure of Horizontal Pyramid Mapping.

HPM有S个刻度，在$s \in 1, 2, ..., S$，SP提取出的特征图在height维度被划分为2^{s-1}个条带，一共有$\sum_{s=1}^{S} 2^{s-1}$个条带。然后一个Global Pooling作用于3-D的条带提取1-D的特征。对于一个条带$z_{s,t}$，其中$t \in 1, 2, ..., 2^{s-1}$表示条带在刻度中的索引，Global Pooling的方程为：$f'_{s,t} = maxpool(z_{s,t}) + avgpool(z_{s,t})$，其中$maxpool$是Global Max Poling，$avgpool$是Global Average Pooling。最后一步是用FCs来将特征f'映射到一个可描述的空间，每一种条带使用一种全连接。

卷积网络的不同层具有不同的接收场。 层越深，接收场将越大。 因此，浅层特征图中的像素集中在局部和细粒度的信息上，而较深层特征图中的像素集中在更全局和粗粒度的信息上。 通过在不同层上应用SP提取的集合级别特征具有类比属性，为了收集各种级别的集合信息，提出了多层全局管道（MGP）。它具有与主管道中的卷积网络类似的结构，并且将在不同层中提取的集级别特征添加到MGP。 由MGP生成的最终特征图也将由HPM映射到$\sum_{s=1}^{S} s^{s-1}$特征中。

最后的损失函数使用的是triplet loss

算法实验结果如下：

Table 1: Averaged rank-1 accuracies on **CASIA-B** under three different experimental settings, excluding identical-view cases.

Gallery NM#1-4			0°-180°										mean
Probe			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
ST (24)	NM#5-6	ViDP (Hu et al. 2013)	–	–	–	59.1	–	50.2	–	57.5	–	–	–	–
		CMCC (Kusakunniran et al. 2014)	46.3	–	–	52.4	–	48.3	–	56.9	–	–	–	–
		CNN-LB (Wu et al. 2017)	54.8	–	–	77.8	–	64.9	–	76.1	–	–	–	–
		GaitSet(ours)	**64.6**	**83.3**	**90.4**	**86.5**	**80.2**	**75.5**	**80.3**	**86.0**	**87.1**	**81.4**	**59.6**	**79.5**
	BG#1-2	GaitSet(ours)	55.8	70.5	76.9	75.5	69.7	63.4	68.0	75.8	76.2	70.7	52.5	68.6
	CL#1-2	GaitSet(ours)	29.4	43.1	49.5	48.7	42.3	40.3	44.9	47.4	43.0	35.7	25.6	40.9
MT (62)	NM#5-6	AE (Yu et al. 2017b)	49.3	61.5	64.4	63.6	63.7	58.1	59.9	66.5	64.8	56.9	44.0	59.3
		MGAN (He et al. 2019)	54.9	65.9	72.1	74.8	71.1	65.7	70.0	75.6	76.2	68.6	53.8	68.1
		GaitSet(ours)	**86.8**	**95.2**	**98.0**	**94.5**	**91.5**	**89.1**	**91.1**	**95.0**	**97.4**	**93.7**	**80.2**	**92.0**
	BG#1-2	AE (Yu et al. 2017b)	29.8	37.7	39.2	40.5	43.8	37.5	43.0	42.7	36.3	30.6	28.5	37.2
		MGAN (He et al. 2019)	48.5	58.5	59.7	58.0	53.7	49.8	54.0	61.3	59.5	55.9	43.1	54.7
		GaitSet(ours)	**79.9**	**89.8**	**91.2**	**86.7**	**81.6**	**76.7**	**81.0**	**88.2**	**90.3**	**88.5**	**73.0**	**84.3**
	CL#1-2	AE (Yu et al. 2017b)	18.7	21.0	25.0	25.1	25.0	26.3	28.7	30.0	23.6	23.4	19.0	24.2
		MGAN (He et al. 2019)	23.1	34.5	36.3	33.3	32.9	32.7	34.2	37.6	33.7	26.7	21.0	31.5
		GaitSet(ours)	**52.0**	**66.0**	**72.8**	**69.3**	**63.1**	**61.2**	**63.5**	**66.5**	**67.5**	**60.0**	**45.9**	**62.5**
LT (74)	NM#5-6	CNN-3D (Wu et al. 2017)	87.1	93.2	97.0	94.6	90.2	88.3	91.1	93.8	96.5	96.0	85.7	92.1
		CNN-Ensemble (Wu et al. 2017)	88.7	95.1	98.2	96.4	**94.1**	91.5	93.9	97.5	98.4	95.8	85.6	94.1
		GaitSet(ours)	**90.8**	**97.9**	**99.4**	**96.9**	93.6	**91.7**	**95.0**	**97.8**	**98.9**	**96.8**	**85.8**	**95.0**
	BG#1-2	CNN-LB (Wu et al. 2017)	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
		GaitSet(ours)	**83.8**	**91.2**	**91.8**	**88.8**	**83.3**	**81.0**	**84.1**	**90.0**	**92.2**	**94.4**	**79.0**	**87.2**
	CL#1-2	CNN-LB (Wu et al. 2017)	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
		GaitSet(ours)	**61.4**	**75.4**	**80.7**	**77.3**	**72.1**	**70.1**	**71.5**	**73.5**	**73.5**	**68.4**	**50.0**	**70.4**