On Learning Disentangled Representations for Gait Recognition Notes

author:Ziyuan Zhang, Luan Tran, Feng Liu, Member, IEEE, and Xiaoming Liu, Member, IEEE

摘要：

当前主要步态识别方法采用人形剪影或者链接类型的人体模型作为步态特征。当面临着装、携带物品或者角度的变化效果步态理想。本文提出了一种自动编码的框架GaitNet来从RGB图像中分隔出外观特征、典范特征（类似于GEI的静态特征）以及姿态特征（使用LSTM，作为动态特征）。同时，文章结合扫了Frontal-View Gait(FVG)这个关注于正面的数据集。

正文

Symbol	Dim.	Notation
s	scalar	Index of subject
c	scalar	Condition
t	scalar	Time step in a video
n	scalar	Number of frames in a video
$\mathbf{X}^c$	matrices	Gait video under condition c
$\mathbf{x}^{c,t}$	matrix	Frame t of video $\mathbf{X}^c$
$\hat{\mathbf{x}}$	matrix	Reconstructed frame via $\mathcal{D}$
$\mathcal{E}$	-	Encoder network
$\mathcal{D}$	-	Decoder network
C^{sg}	-	Classifier for $\mathbf{f}_c$
C^{dg}	-	Classifier for $\mathbf{f}_{\text{dyn-gait}}$
$\mathbf{f}_p$	64×1	Pose feature
$\mathbf{f}_c$	128×1	Canonical feature
$\mathbf{f}_a$	128×1	Appearance feature
$\mathbf{f}_{\text{dyn-gait}}$	256×1	Dynamic gait feature
$\mathbf{f}_{\text{sta-gait}}$	128×1	Static gait feature
$\mathbf{h}^t$	128×1	The output of LSTM at step t
$Ł_{\text{xrecon}}$	-	Reconstruction loss
$Ł_{\text{pose-sim}}$	-	Pose similarity loss
$Ł_{\text{cano-sim}}$	-	Canonical similarity loss
$Ł_{\text{id-inc-avg}}$	-	Incremental identity loss

符号列表

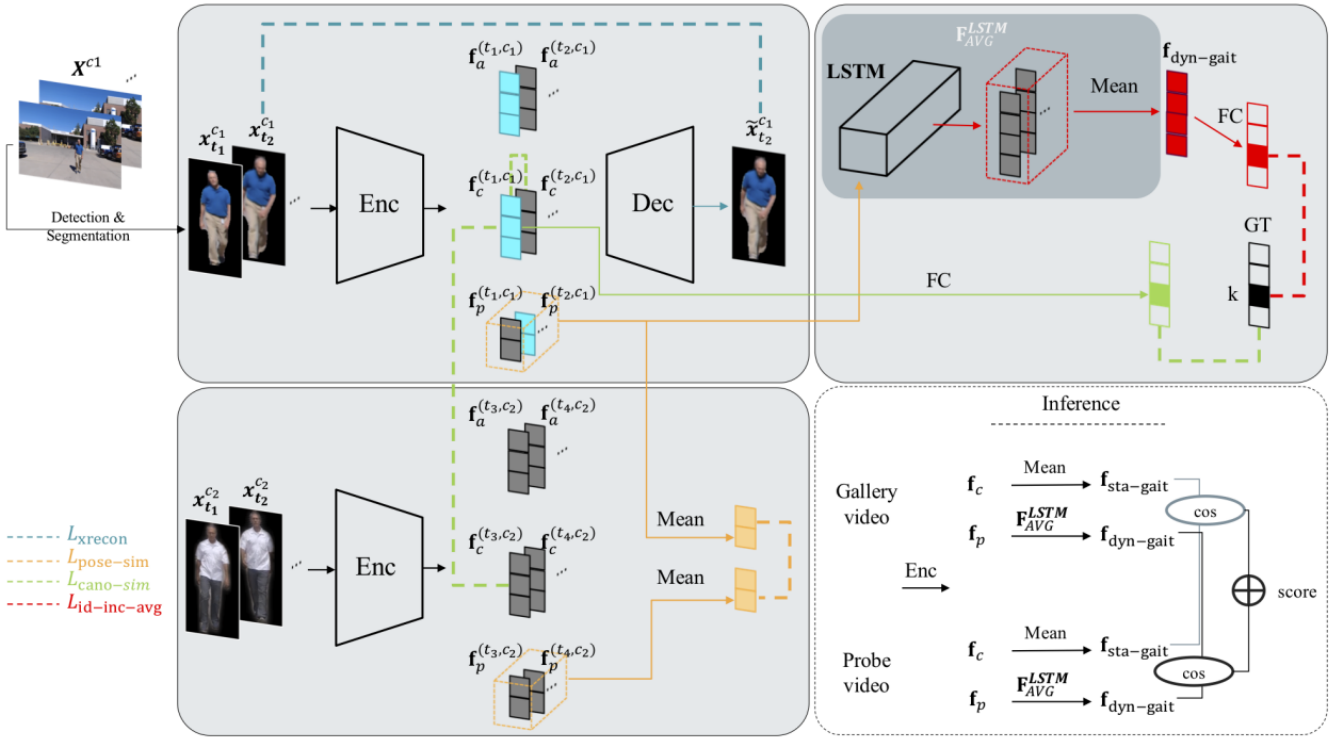

Fig. 3: The overall architecture of proposed GaitNet. The bottom right block indicates the inference process, while the remaining illustrates the training process with the four color-coded loss functions.

算法结构

文章做了特征分解以解决之前提出的一系列问题。

pose feature：外观特征描述对象的衣服

appereance feature：姿势特征描述了身体部位的位置，其随时间的动态变化是步态的核心要素

canonical feature：描述身材的特征（矩形，三角形，倒三角形和沙漏形身材），手臂长度、躯干vs腿长的比例等

算法流程

输入GaitNet的是一段视频序列，使用目标检测与分割算法（文中使用Mask R-CNN）提取人物，设计三种损失函数，编码后的数据用于区分pose,canonical,appearance

特征分解

编码算法E：$f_a, f_c, f_p = E(x)$

解码算法D：$\hat{x} = D(f_a, f_c, f_p)$

编码解码算法使用的CNN

$\mathcal{E}$			$\mathcal{D}$		
Layers	Filters/Stride	Output Size	Layers	Filters/Stride	Output Size
Conv1	3x3/1	64x32x64	FC	-	4x2x512
MaxPool1	3x3/2	32x16x64	FCConv1	3x3/2	8x4x256
Conv2	3x3/1	32x16x256	FCConv2	3x3/2	16x8x128
MaxPool2	3x3/2	16x8x256	FCConv3	3x3/2	32x16x64
Conv3	3x3/2	16x8x512	FCConv4	3x3/2	32x16x3
(Conv4	3x3/2	16x8x512)*			
MaxPool3	3x3/2	4x2x512			
FC	-	320			

交叉重建损失(Cross Reconstruction Loss)：$\hat{x}$与原始输入的x接近。文章的cross reconstruction loss使用帧t_1的appearance特征${f_a}^{t_1}$和canonocal特征${f_c}^{t_1}$以及帧t_2的姿势特${f_p}^{t_2}$，

$$L_{xrecon} = ||D(f_a^{t_1}，\ f_c^{t_1}，\ f_p^{t_2}) - x^{t_2}||$$

姿势相似度损失(Pose Similarity Loss)：类似于GEI，同一个人的两段视频c_1长n_1，c_2长n_2，

$$L_{pose-sim} = ||\frac{1}{n_1}\sum_{t=1}^{n_1} f_p^{(t,c_2)}||_2^2$$

规范一致性损失(Canonical Consistency Loss)：

$$L_{cano-cons} = \frac{1}{n_1^2}\sum_{i\neq j}||f_c^{(t_i,c_1)} - f_c^{(t_j,c_1)}||_2^2 + \frac{1}{n_1}\sum_i||f_c^{(t_i,c_1)} - f_c^{(t_j,c_2)}||_2^2 + \frac{1}{n_1}\sum_i -log(C_k^{sg}(f_c^{(t_1,c_1)}))$$

步态特征学习与聚合

通过canonical特征聚合得到静态步态特征

$$f_{sta-gait} = \frac{1}{n}\sum_{t=1}^{n} f_c^t$$

通过聚合pose特征获得动态步态特征

使用LSTM探索pose特征的时间信息，比如：受试者身体部位的轨迹如何随时间变化

LSTM的输出送到了一个C^{dg}分类器中（文中使用线性分类器）

设h^t为LSTM在时间t的输出

$$h^t = LSTM(f_p^1, f_p^2, ..., f_p^t)$$

一个简单的识别方法是在最终时间步骤的LSTM输出之上添加分类损失

$$L_{id-single} = -log(C_k^{dg}(h^n))$$

使用平均特征进行识别（Identification with Averaged Feature）

为了减少最后一个时间数据对LSTM输出的影响，文中使用平均的LASTM输出作为步态特征来识别

$$f_{dyn-gait}^t = \frac{1}{t}\sum_{s=1}^{t} h^s$$

识别损失函数为：

$$L_{id-avg} = -log(C_k^{dg}(f_{dyn-gait}^n)) = -log(C_k^{dg})(\frac{1}{n}\sum_{s=1}^{n} h^s)$$

增量身份损失函数（Incremental Identity Loss）
预计LSTM会了解到，视频序列越长，它处理的步行信息越多，因此可以更加有效地识别出对象。 文中使用每个时间步的所有中间输出（按$w_t = t^2$)加权）：

$$L_{id-inc-avg} = \frac{1}{\sum_{t=1}^{n} w_t}\sum_{t=1}^{n} -w_t log(C_k^{dg}(\frac{1}{t}\sum_{s=1}^{t} h^s))$$

最终的损失函数：

$$L = L_{id-inc-avg} + \lambda_r L_{xrecon} + \lambda_d L_{pose-sim} + \lambda_s L_{cano-sim}$$

模型推论

g代表图库（gallery），p代表需探测目标（probe）

$$Score = (1-\alpha) * cos(f_{sta-gait}^g, f_{sta-gait}^p) + \alpha * cos(f_{dyn-gait}^g, f_{dyn-gait}^p)$$

结论

TABLE 7: Comparison on CASIA-B with cross view and conditions. Three models are trained for NM-NM, NM-BG, NM-CL. Average accuracies are calculated excluding probe viewing angles.

Gallery NM #1-4	0°-180° (exclude identical viewing angle)											
Probe NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
ViDP [65]	-	-	-	64.2	-	60.4	-	65.0	-	-	-	-
LB [11]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
3D MT network [11]	87.1	93.2	97.0	94.6	90.2	88.3	91.1	93.8	96.5	96.0	85.7	92.1
J-CNN [34]	87.2	93.2	96.3	95.9	91.6	86.5	89.8	93.8	95.1	93.0	80.8	91.2
GaitNet-pre [29]	91.2	92.0	90.5	95.6	86.9	92.6	93.5	96.0	90.9	88.8	89	91.6
GaitNet	**93.1**	**92.6**	**90.8**	**92.4**	**87.6**	**95.1**	**94.2**	**95.8**	**92.6**	**90.4**	**90.2**	**92.3**
Probe BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
LB-subGEI [11]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
J-CNN [34]	73.1	78.1	83.1	81.6	71.6	65.5	71.0	80.7	79.1	78.6	68.0	75.0
GaitNet-pre [29]	83.0	87.8	88.3	93.3	82.6	74.8	89.5	91.0	86.1	81.2	85.6	85.7
GaitNet	**88.8**	**88.7**	**88.7**	**94.3**	**85.4**	**92.7**	**91.1**	**92.6**	**84.9**	**84.4**	**86.7**	**88.9**
Probe CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
LB-subGEI [11]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	53.98
J-CNN [34]	46.1	58.4	64.4	64.2	55.5	50.5	54.7	55.8	53.3	51.3	39.9	54.01
GaitNet-pre [29]	42.1	58.2	65.1	70.7	68.0	70.6	65.3	69.4	51.5	50.1	36.6	58.9
GaitNet	**50.1**	**60.7**	**72.4**	**72.1**	**74.6**	**78.4**	**70.3**	**68.2**	**53.5**	**44.1**	**40.8**	**62.3**

TABLE 8: Recognition accuracy cross views under NM on CASIA-B dataset. One single GaitNet model is trained for all the viewing angles.

Methods	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°	Average
CPM [33]	13	14	17	27	62	65	22	20	15	10	24.1
GEI-SVR [60]	16	22	35	63	95	95	65	38	20	13	42.0
CMCC [66]	18	24	41	66	96	95	68	41	21	13	43.9
ViDP [65]	8	12	45	80	**100**	100	81	50	15	8	45.4
STIP+NN [61]	-	-	-	-	84.0	86.4	-	-	-	-	-
LB [11]	18	36	67.5	93	99.5	99.5	92	66	36	18	56.9
L-CRF [33]	38	75	68	93	98	99	93	67	76	39	67.8
GaitNet-pre [29]	68	74	88	91	99	98	84	75	76	65	81.8
GaitNet	**82**	**83**	86	91	93	98	**92**	**90**	**79**	**79**	**87.3**