文章编号：2096-1472(2023)-06-20-04
DOI:10.19644/j.cnki.issn2096-1472.2023.006.005

基于网格划分骨骼的行为预测

沈江霖，魏　丹，王子阳

(上海工程技术大学机械与汽车工程学院，上海　201620)
✉ goatlmljrf@163.com; weidan@sues.edu.cn; wangziyangwilliam@163.com

摘　要：行为预测会面临行人部位遮挡、背景干扰、摄像机视角不同、行人姿态不同、外观差异过大及行人动作信息提取难度过大等问题。文章提出一种新的基于网格划分骨骼的行为预测方法，首先使用自下而上的方法提取行人的骨骼信息，通过学习人体关节点的距离度量特征和角度度量特征提取行人的行为特征。然后对关节点分别对比前后帧的行为特征，判断下一帧单个关节点运动类型发生的概率，通过对下一帧关节点运动类型的加权判断下一帧行人的动作。所提方法预测人体右脚关节点向上运动的概率为92.3%。

关键词：行为预测；自下而上；关节点；距离度量
中图分类号：TP181　　文献标识码：A

Action Prediction Method Based on Grid Partition Skeleton

SHEN Jianglin, WEI Dan, WANG Ziyang

(*School of Mechanical and Automobile Engineering, Shanghai University of Engineering Science, Shanghai* 201620, *China*)
✉ goatlmljrf@163.com; weidan@sues.edu.cn; wangziyangwilliam@163.com

Abstract: Action prediction faces problems such as pedestrian parts occlusion, background interference, different camera angles, different pedestrian posture, large appearance difference, and difficulty in extracting pedestrian motion information. This paper proposes a new action prediction method based on grid partitioning skeleton. Firstly, the bottom-up method is used to extract the bone information of pedestrians, and then the action characteristics of pedestrians are extracted by learning the distance measurement features and angle measurement features of human joints. By comparing human joints action characteristics of two consecutive frames, the probability of the movement type of a single joint in the next frame is determined, and the movement of pedestrians in the next frame is determined by weighting the movement type of the next frame. The probability of predicting the upward movement of the right foot joint by the proposed method is 92.3%.

Keywords: action prediction; bottom-up; key point; distance measure

1　引言(Introduction)

在无人驾驶中，通过对行人下一步动作的预测可以提前完成汽车加速或者减速的决策，同时可以减小事故发生的概率[1]。如果不能准确的预测行人的下一步动作，人车系统安全将无从谈起。CHEN等[2]提出了一种用于动作预测的循环语义保留生成方法，开发了一个生成体系结构补充骨架序列用于预测动作，该方法未考虑行人部位遮挡及背景干扰问题。LI等[3]研究了基于骨架数据的动作预测，提出了一种基于对抗学习的自适应图卷积网络，学习局部序列中潜在的全局信息，该网络对行人姿态变化和外观差异等因素不具有鲁棒性。针对上述问题，本文提出一种新的基于网格划分骨骼的行为预测方法，通过对行人进行网格划分并提取相应关节点的行为特征，对比前后帧的行为特征，进而判断行人的运动方向及运动速度。

2　行为预测(Action prediction)

基于网格划分骨骼的行为预测方法主要分为两个部分：行人网格划分和行人骨骼行为特征提取。具体来说，就是将行人进行网格划分和对行人骨骼进行估计，提取出网格特征、

关节点角度及骨骼特征，将三者结合形成行人的行为动作特征，通过前后帧的对比，计算下一帧行人动作发生的概率。基于网格划分骨骼的行为预测方法的流程如图1所示。

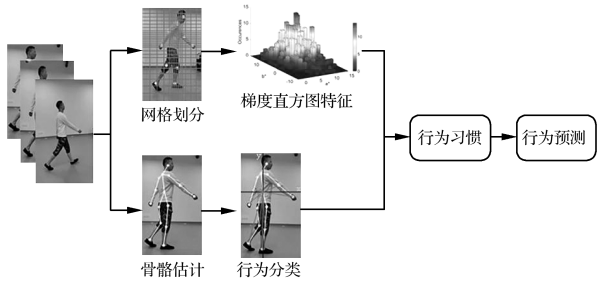

图1　基于网格划分骨骼的行为预测方法的流程图

Fig. 1 Flow chart of action prediction method based on grid partition of skeleton

给定一个测试视频，初始化其定位算法，在单帧中使用自下而上的骨骼估计，并使用之前帧的多个时空约束细化位姿。本文将测试视频帧进行密集网格划分，利用每个划分网格内计算的特征，学习1个基于划分网格的外观模型，该模型通过训练1个在每个位姿边界框内的网格特征作为前景，其余划分的网格特征作为背景的判别分类器，区分前景和背景。同时，由于骨骼关节点角度和前一帧骨骼关节点角度是一致的，基于此可以计算当前时间步长姿态假设的条件概率。将每个时间步长的概率结果结合后，通过在关节点位置、行人外观及姿势比例上施加一致性改进姿势。

一旦在当前的时间步长中估计并改进了姿势，就会更新基于网格特征的外观模型，以避免出现视觉漂移。因此，骨骼估计不但提供了初始化具有判别力的外观模型，而且可以包含任何行人全身或多个关节点的交互或执行操作。

本文采取文献[4]中介绍的自下而上的方法进行行人骨架估计，行人骨架估计结果如图2左部分和图3左部分所示。行人骨架估计完成后，本文将骨架特征单独提取出来，如图2右部分和图3右部分所示，并规定了8个关节点作为行为角度特征的基点，分别为左右肘关节点(A,a,C,c)、左右手关节点(B,b,D,d)、左右膝关节点(E,e,I,i)、左右脚关节点(F,f,J,j)。本文通过网格划分得到的网格中心 o 作为二维坐标系的原点，以 o 为原点学习8个关节点到原点的距离度量特征和角度度量特征。通过与前一帧的距离度量特征和角度度量特征进行对比，判断各个关节点的变化趋势。

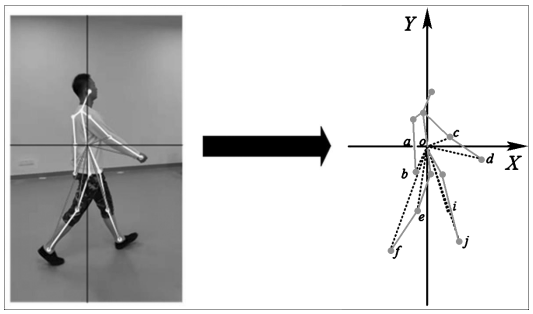

图2　前一帧行人骨骼估计及其角度特征

Fig. 2 Pedestrian skeleton estimation and its angle characteristics in the previous frame

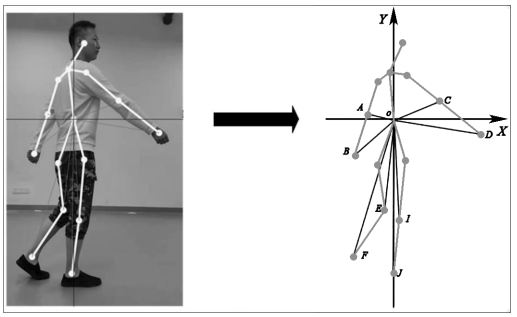

图3　后一帧行人骨骼估计及其角度特征

Fig. 3 Post-frame pedestrian skeleton estimation and its angular characteristics

当行人行走时，o 点与头部、肩部和胯部相应关节点的距离度量和角度度量变化不大，不能体现行人的姿态变化，更无法通过 o 点与头部、肩部和胯部相应关节点的距离度量和角度度量预测行人下一帧的动作变化。与头部、肩部和胯部相应的关节点不同的是，行人行走时的左右肘关节点、左右手关节点、左右膝关节点、左右脚关节点与 o 点的距离度量和角度度量变化幅度很大；而且行人行走时，手臂的摆幅、跨步的距离都是行人行走习惯的表现，通过对左右肘关节点、左右手关节点、左右膝关节点、左右脚关节点与 o 点的距离度量和角度度量的学习，可以得到行人的行走习惯，从而判断下一帧行人的动作变化。

不同帧捕捉到行人的左右肘关节点、左右手关节点、左右膝关节点和左右脚关节点的距离度量特征和角度度量特征不同，通过对比前后帧的距离度量特征和角度度量特征计算出下一帧关节点的运动类型发生概率。本文利用公式(1)计算出关节点的距离度量特征 d_{oi}：

$$d_{oi}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} \tag{1}$$

本文定义的行为特征包括距离度量特征和角度度量特征，由于关节点与原点 o 的角度不易计算，因此本文利用关节点角度的正弦值表示关节点的角度。本文利用公式(2)计算角度度量特征 θ_i：

$$\theta_i=\arcsin\left(\frac{x_i}{d_{oi}}\right) \tag{2}$$

行人行为特征的距离度量特征 d_{oi} 和角度度量特征 θ_i 都是基于相同关节点计算得出，两者具有相关性。本文利用公式(3)计算出两者的关联度 τ_i：

$$\tau_i=\frac{\min\{\sin\theta_i\}\cdot\min\{d_{oi}\}}{1+\rho\max\{\sin\theta_i\}\cdot\max\{d_{oi}\}}+\frac{\rho\max\{\sin\theta_i\}\cdot\max\{d_{oi}\}}{1+\rho\max\{\sin\theta_i\}\cdot\max\{d_{oi}\}} \tag{3}$$

其中，ρ 为分辨系数，一般 $\rho=0.5$。通过关联度 τ_i 将两个特征关联后得到行人的行为特征 T_i，行为特征 T_i 可以用公式(4)表示：

$$T_i=\sin\theta_i+\tau_i d_{oi} \tag{4}$$

计算出行人行为特征后，利用公式(5)计算下一帧该关节点的运动特征出现概率：

$$P^m_{i_{t+1}}=\frac{\exp(T^m_i)}{\sum_{m=1}^{M}\exp(T^m_i)} \tag{5}$$

其中，$P^{m}_{i_{t+1}}$ 表示 $t+1$ 时刻关节点运动类型为 m 的概率，m 为关节点运动类型，M 为关节点运动类型的总数。

利用上述公式计算出 8 个关节点下一帧运动类型的概率，然后结合 8 个关节点下一帧的运动类型判断下一帧行人动作。为了更准确地判断行人运动，本文将行人划分为无数密集网格，结果如图 4 所示，通过提取划分网格的特征，对比前后帧的 8 个关节点的网格特征变化，判断行人前进或者后退，以及行动加速、匀速或者减速。以上判断都是基于 MATLAB 代码实现的。

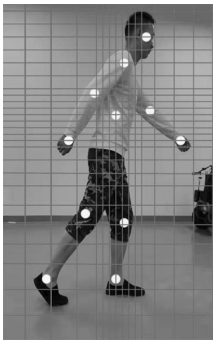

图 4 行人网格划分结果

Fig. 4 Pedestrian grid partition results

3 实验与结果分析(Experiments and analysis of results)

3.1 数据集

在 3 个行为动作预测数据集 sub-JHMDB、UCF-Sports 和 MSR-II 上对本文所提方法进行验证。

sub-JHMDB 数据集[5]的每帧都可以看到所有人体关节，共包含 316 个视频、12 个动作类。考虑到行人关节的复杂变化，每一帧行人的识别和定位变得极具挑战性，故在 JHMDB 数据集基础上，使用 sub-JHMDB 数据集。

UCF-Sports 数据集[6-7]包含 150 个视频和 10 个动作类。使用文献[8]提出的方法对本文所提方法进行评估。

MSR-II 数据集[9]包含 54 个未修剪的视频和 3 个动作类。本文采用跨数据集估值，并使用 KTH 数据集对 MSR-II 数据集进行训练和测试。在数据集 iDTFs 使用大小为 $K=1\ 000$ 的码数训练支持向量机。本文使用精确回忆曲线和最先进的离线方法进行定量比较，同时为了与其他数据集保持一致，本文也使用受试者工作特征曲线(Receiver Operating Characteristic，ROC)和 ROC 曲线下面区域的面积(Area Under Curve，AUC)报告结果。

3.2 实验设置

受早期行为识别和预测的启发，本文将行为和交互的观察比值作为性能量化指标。该评价指标以不同的观测视频/动作比值(0，0.1，0.2，…，1)采样，进行视频的定位和预测。预测任务的准确性类似于分类和识别，对于未修剪的视频，预测精度的评估难度较大。为此，首先将预测结果作为观测比值函数的真实视频行动，该操作相当于修剪情况。其次使用动作/交互的平均持续时间，通过在每个视频中每 5 帧滑动一个窗口提取时间上重叠的剪辑，其中，一些包含真实动作，另一些则代表未修剪视频的背景部分。最后将计算预测精度作为函数的观测比值。这种方法可以捕捉误报，并为未修剪的视频提供更全面的评估。

3.3 实验结果

在实验部分，本文以连续 3 帧为例，通过学习前两帧的行为特征估计下一帧的 8 个关节点的运动类型发生的概率，根据 8 个关节点运动类型的加权后得到行人下一帧的运动类型。

通过上述方法可以计算出 8 个关节点的运动类型发生的概率，结果如表 1 所示。从表 1 中可以推测出，行人的左臂在下一帧向上运动，行人的右臂在下一帧向下运动，行人的左腿在下一帧向下运动，行人的右腿在下一帧向上运动。根据划分网格特征判断行人向前匀速运动，其中左右手臂和左右腿的运动方向也是向前运动。

表 1 关节点运动类型及概率

Tab.1 Type and probability of joint movement

关节点	运动类型	概率/%
A	向上	78.7
B	向上	84.4
C	向下	69.9
D	向下	87.9
E	向下	70.5
F	向下	85.2
I	向上	88.7
J	向上	92.3

研究人员在 3 个行人动作数据集 sub-JHMDB、UCF-Sports、MSR-II 上验证在不同观察比值情况下不同重叠率的动作预测准确率。具体实验结果如图 5 至图 7 所示。其中，当动作观察比值为 0.4、重叠率为 30%时，在 sub-JHMDB 数据集上的准确率能达到 56%；当动作观察比值为 0.2、重叠率为 10%时，在 UCF-Sports 数据集上的准确率能达到 54%；当动作观察比值为 0.1、重叠率为 10%时，在 MSR-Ⅱ数据集上的准确率能达到 46%，不同数据集上最大准确率和动作观察比值见表 2。

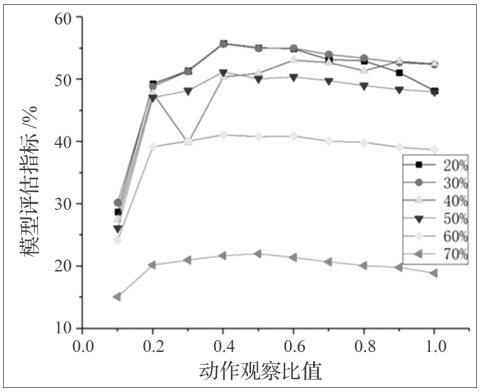

图 5 不同观察比值在 sub-JHMDB 数据集上的准确率

Fig. 5 The accuracy of different observation ratios on sub-JHMDB dataset

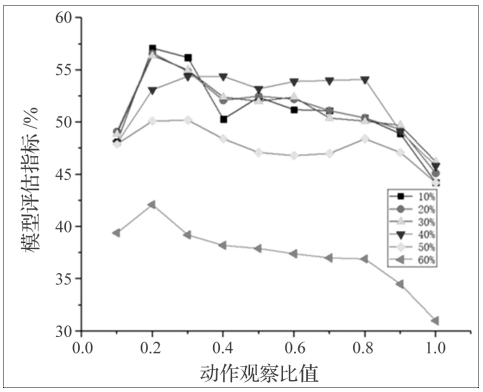

图 6　不同观察比值在 UCF-Sports 数据集上的准确率

Fig. 6 The accuracy of different observation ratios on UCF-Sports dataset

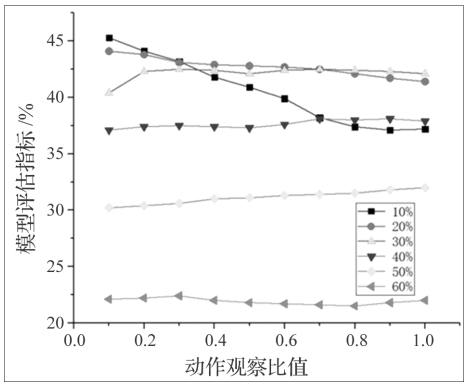

图 7　不同观察比值在 MSR-II 数据集上的准确率

Fig. 7 The accuracy of different observation ratios on MSR-II dataset

表2　不同数据集上最大准确率和动作观察比值

Tab.2 Maximum accuracy and action observation ratios on different datasets

数据集	动作观察比值	准确率/%
sub-JHMDB	0.4	56
UCF-Sports	0.2	54
MSR-II	0.1	46

本研究以累积的方式计算 AUC，50%的准确率意味着从一开始就定位一个动作，直到观察到视频的一半。这可以深入了解本文所提模型性能是如何随着时间或测试视频中观察到的比值的变化而变化的。从图 5 至图 7 可以看出，在视频开始播放时，定位一个动作是很有挑战性的，因为算法并没有观察到足够的判别运动用于区分不同的动作。此外，本研究首先从行人骨架姿态中学习外观模型，并随着时间的推移不断进行改进和细化。这提高了基于划分网格特征的外观可信度，进而提高了预测的准确性，稳定了 AUC。

4　结论(Conclusion)

本文提出一种新的基于网格划分骨骼的行为预测方法，该方法主要分为两个部分，首先使用自下而上的方法提取行人的骨骼信息，规定左右肘、左右手、左右膝以及左右脚 8 个关节点作为提取行为习惯的关节点，并且通过学习 8 个关节点的距离度量特征和角度度量特征提取行人的行为特征。对 8 个关节点分别对比前后帧的行为特征，判断下一帧单个关节点运动类型发生的概率，通过对下一帧 8 个关节点运动类型的加权判断下一帧行人的动作。其次为了更好地评估行人下一帧的动作，本文通过对行人进行网格划分并提取相应关节点的划分网格特征，对比前后帧的划分网格特征，判断行人的运动方向及运动速度。实验结果验证了该方法的有效性。

参考文献(References)

[1] 莫晨,邵洁.基于自注意力的多模态 LSTM 的动作预测[J].计算机工程与设计,2022,43(4):1083-1088.

[2] CHEN L, LU J, SONG Z, et al. Recurrent semantic preserving generation for action prediction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1):231-245.

[3] LI G, LI N, CHANG F, et al. Adaptive graph convolutional network with adversarial learning for skeleton-based action prediction[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(3):1258-1269.

[4] CAO Z, SIMON T, WEI S, et al. Realtime multi-Person 2D pose estimation using part affinity fields[C]// Institute of Electrical and Electronics Engineers. IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE, 2017:1302-1310.

[5] JHUANG H, GALL J, ZUFFI S, et al. Towards understanding action recognition[C]// Institute of Electrical and Electronics Engineers. IEEE International Conference on Computer Vision. Los Alamitos: IEEE, 2013:3192-3199.

[6] ZHENG L, HUANG Y, LU H, et al. Pose invariant embedding for deep person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(9):4500-4509.

[7] BAI S, BAI X, TIAN Q. Scalable person re-identification on supervised smoothed manifold[C]// Institute of Electrical and Electronics Engineers. IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE, 2017:3356-3365.

[8] LAN T, WANG Y, MORI G. Discriminative figure-centric models for joint action localization and recognition[C]// Institute of Electrical and Electronics Engineers. IEEE International Conference on Computer Vision. Los Alamitos: IEEE, 2011:2003-2010.

[9] YUAN J, LIU Z, WU Y. Discriminative video pattern search for efficient action detection[J]. IEEE Transaction on Pattern Anally Machine, 2011, 33(9):1728-1743.

作者简介：

沈江霖(1997-)，男，硕士生. 研究领域：行人重识别，图像处理.

魏　丹(1982-)，女，博士，副教授. 研究领域：行人重识别，图像处理.

王子阳(1991-)，男，硕士. 研究领域：行人重识别，图像处理.