基于多源特征融合的行人穿越行为预测

侯林鹏[1],杨超宇[2]
1. 安徽理工大学 计算机科学与工程学院,安徽 淮南 232001
2. 安徽理工大学 人工智能学院,安徽 淮南 232001

摘 要:目的 在智能驾驶领域,准确预测行人穿越行为对于确保车辆和行人安全至关重要。**方法** 设计了一种结合多种计算机视觉技术的行人穿越行为预测模型,该模型通过分析行人的位置、姿态、动作及环境特征来准确判断行人意图。为了增强模型对不同距离行人的感知能力,采用了不同尺度放大的预处理和数据滤波平滑的后处理技术。提出了先条件后预测(Predict after Condition,PAC)两阶段方法,以实现更为有效的行人穿越预测。**结果** 基于 JAAD 数据集的测试结果表明:所提模型平均精度达 89.31%,相较于传统单阶段方法提升了 8.76%。**结论** 特征重要度分析进一步表明:加入路面面积特征后,预测准确率从 68.43%显著提升至 85.06%,强调了行人位置与路面轮廓关系在行人穿越行为研究中的重要性。对降低人车碰撞事故,提高智能驾驶车辆的安全性具有重要意义。

关键词:行人穿越行为预测;多特征融合;行人行为;智能驾驶
中图分类号:U471.15 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2026.0002.011

Pedestrian Crossing Behavior Prediction Based on Multi-Source Feature Fusion

HOU Linpeng[1], YANG Chaoyu[2]
1. School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan 232001, Anhui, China
2. School of Artificial Intelligence, Anhui University of Science & Technology, Huainan 232001, Anhui, China

Abstract: Objective Accurately predicting pedestrian crossing behavior is crucial for ensuring the safety of both vehicles and pedestrians in the field of intelligent driving. **Methods** A pedestrian crossing behavior prediction model was designed, which combined various computer vision technologies. The model accurately assessed pedestrian intentions by analyzing their positions, postures, actions, and environmental features. To enhance the model's perception capability for pedestrians at different distances, preprocessing with multi-scale magnification and post-processing for data filtering and smoothing were employed. A two-stage method called prediction after condition (PAC) was proposed to achieve more effective prediction of pedestrian crossings. **Results** Testing results based on the JAAD dataset indicated that the proposed model achieved an average accuracy of 89.31%, representing an improvement of 8.76% compared with traditional single-stage methods. **Conclusion** Further analysis of feature importance shows that after incorporating the road surface area feature, the prediction accuracy significantly increases from 68.43% to 85.06%, highlighting the importance of the relationship between pedestrian location and the road profile in the study of pedestrian crossing behavior. This has significant implications for reducing vehicle-pedestrian collision accidents and enhancing the safety of intelligent driving vehicles.

Keywords: pedestrian crossing behavior prediction; multi-feature fusion; pedestrian behavior; intelligent driving

收稿日期:2024-03-19 **修回日期:**2024-06-14 **文章编号:**1672-058X(2026)02-0083-10
基金项目:国家自然科学基金项目(61873004)资助.
作者简介:侯林鹏(1998—),男,硕士研究生,从事计算机视觉研究.
通信作者:杨超宇(1981—),男,博士,教授,从事计算机视觉研究. Email:yangchy@aust.edu.cn.
引用格式:侯林鹏,杨超宇. 基于多源特征融合的行人穿越行为预测[J]. 重庆工商大学学报(自然科学版),2026,43(2):83-92.
HOU Linpeng, YANG Chaoyu. Pedestrian crossing behavior prediction based on multi-source feature fusion[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2026, 43(2): 83-92.

随着人工智能技术的加速发展以及数据和计算能力的不断扩大,高级自动驾驶系统在未来交通体系中的广泛应用正变得日益可行[1]。根据世界卫生组织的统计报告,全球每年大约有135万人死于道路交通事故[2]。行人、骑行人等弱势道路使用者在穿越时需要时刻注意周边来车情况,并做出等待、避让或穿越行为。而对于自动驾驶汽车而言,必须不断与弱势道路使用者交互。因此,如何有效检测并预测行人的穿越意图是关键。

目前,对行人穿越预测的研究主要分为基于行人未来轨迹预测的方法和基于人体姿态估计的方法。其中基于行人未来轨迹预测的方法主要是借助先验历史轨迹信息建模来预测未来轨迹,目前多数的研究大都使用基于RNN的编码器-解码器框架进行预测,较常见的为RNN(Recurrent Neural Network)的变体:LSTM(Long Short-Term Memory)与GRU(Gated Recurrent Unit),Alahi等[3]为了解决行人轨迹预测问题,提出了Social-LSTM网络结构以此捕捉人群中个体之间的交互和影响。Hug等[4]结合了LSTM网络和条件粒子滤波方法,用于增强行人轨迹预测的性能。李文礼等[5]提出一种基于Social-GAN的行人轨迹预测算法SAN-GAN,通过模拟行人的视野域来更准确地预测其未来的位置信息。陈龙[6]结合车辆前视视角下的全局环境、行人姿态等多模态信息,在GRU基础上添加注意力机制以提高模型全局场景上下文交互信息的提取能力。基于人体姿态估计的方法则是考虑以行人二维或三维的骨架在一段时间序列内的微量变化,并以此提取特征预测穿越行为。其常见的方法分为CNN(Convolutional Neural Network)和GCN(Graph Convolutional Network)两种,Rasouli等[7]使用CNN融合行人在场景中的定位信息、行为信息、全局场景信息并预测,但由于未考虑当前行人的局部场景信息,导致预测结果波动较大。Kotseruba等[8]结合时空特征和人车信息,提出了一种基于3D CNN的预测方法。YAN等[9]首次提出使用时空图卷积网络(Spatial Temporal Graph Convolutional Networks,ST-GCN)提取时序骨架的动作特征。Chen等[10]基于图卷积网络处理行人轨迹和穿越行为相关的结构化数据,从行人的运动模式中学习特征并进行预测。胡远志等[11]通过改进的ST-GCN结构提取行人骨架过街特征并取得较好的识别效果,但考虑因素较为单一。

综上所述,轨迹预测方法[3-5]因以相机拍摄的俯瞰图作为数据输入,其虽包含更多交通参与者的全局交汇信息,如行人局部环境、人车交汇轨迹等,却受限于移动相机拍摄的俯视视角环境。基于人体姿态估计的方法[10-11]多以车载前视相机拍摄的视频作为输入源,满足当前智能驾驶的通用环境,但在复杂的交通环境中,行人肢体关键点拟合会受人车距离、相机抖动、目标遮挡等影响造成骨架信息丢失[12]。另外,一些研究[6-7]由于未考虑行人未来轨迹是否与主车存在交汇的趋势,从而使那些并不构成交汇条件的行人也参与了预测过程,当应对多人场景时将严重增加系统的预测噪声和计算代价。

针对以上问题,本文综合轨迹预测方法中的历史人车交汇轨迹、局部环境特征和人体姿态估计方法中的行人过路骨架特征,以车载前视视角环境作为研究背景,设计了一种基于多源特征融合的行人穿越行为预测模型。通过检测、追踪、姿态估计、动作识别、语义分割等多种方法提取交通环境中的多源信息特征,并使用分类器融合这些特征预测得到穿越概率。其主要包括如下两部分内容:

(1) 基于多方法的行人信息获取。以多方法得到时序内行人的位置信息、身体信息、动作信息、环境信息。针对复杂环境下造成的行人骨架信息丢失,采用不同尺度放大的预处理和数据滤波平滑的后处理进行数据增强,提高信息获取性能。

(2) 基于多特征融合的穿越行为预测。通过引入基线建立行人与主车之间的关系,并提取历史信息特征,这些特征通过分类器融合并预测穿越行为。针对行人轨迹是否与主车构成交汇条件提出了先条件后预测的两阶段方法,以消除预测噪声,提高准确性。

1 模型架构设计

本文设计的基于多方法获取行人信息以及融合多源特征预测穿越行为的架构示意图如图1所示,其主要模块概述如下:

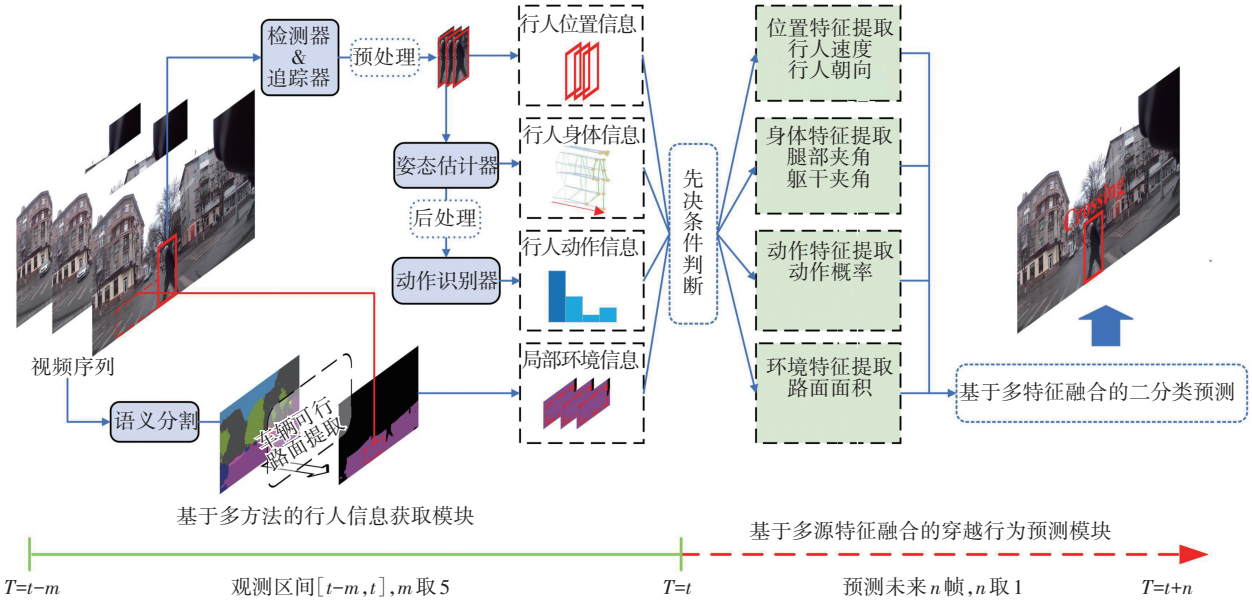


图 1　模型架构

Fig. 1　Model architecture

首先对视频序列中的行人进行检测和追踪,以获取行人的位置信息。然后,利用人体姿态估计器对预处理后的行人进行拟合,得到身体信息。同时,维护一个时序骨架向量,用于记录行人在连续帧中的运动骨架,并将该向量输入给基于骨架的动作识别器中,以获取行人的动作信息。此外,利用语义分割方法对车辆可行路面进行提取,得到环境信息。经多方法得到行人信息序列向量 $\boldsymbol{I}$ 作为预测模块的输入,即:

$$\boldsymbol{I}=(\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_{k-1},\boldsymbol{I}_k) \tag{1}$$

式(1)中,$\boldsymbol{I}_k=(\boldsymbol{S}_k,\boldsymbol{P}_k,\boldsymbol{A}_k,\boldsymbol{E}_k)$;$\boldsymbol{S}_k$、$\boldsymbol{P}_k$、$\boldsymbol{A}_k$、$\boldsymbol{E}_k$ 分别为行人在 k 时刻的位置、身体、动作、环境信息向量。

在行人信息序列中,根据历史信息提取特征。所提取的行人特征序列向量 $\hat{\boldsymbol{I}}$ 表示为

$$\hat{\boldsymbol{I}}=(\hat{\boldsymbol{I}}_1,\hat{\boldsymbol{I}}_2,\cdots,\hat{\boldsymbol{I}}_{k-1},\hat{\boldsymbol{I}}_k) \tag{2}$$

式(2)中,$\hat{\boldsymbol{I}}_k=(\hat{\boldsymbol{S}}_k,\hat{\boldsymbol{P}}_k,\hat{\boldsymbol{A}}_k,\hat{\boldsymbol{E}}_k)$;$\hat{\boldsymbol{S}}_k$、$\hat{\boldsymbol{P}}_k$、$\hat{\boldsymbol{A}}_k$、$\hat{\boldsymbol{E}}_k$ 分别为行人在 k 时刻的位置、身体、动作、环境特征向量。

$\hat{\boldsymbol{I}}$ 经先决条件筛选得到当前满足交汇条件的行人特征集合,再对该特征集合由分类器融合得到穿越预测的概率分布 $\eta=(\eta_{\mathrm{NC}},\eta_{\mathrm{C}})$,其中 η_{NC}、η_{C} 分别为行人未穿越和穿越的概率。

2　基于多方法的行人信息获取

行人的多源信息通过检测器、追踪器、姿态估计器、动作识别器、语义分割等方法获取。

2.1　位置信息获取

使用 YOLOv7[13] 作为检测器,它以其出色的检测精度、实时检测等特点在计算机视觉领域广受关注。某一帧的行人目标通过检测得到目标矩形边界框,其中边界框参数包括矩形左上角点横纵坐标 (x,y) 与矩形的宽高 (w,h)。使用 ByteTrack[14] 作为追踪器,利用连续帧之间的时序信息,对目标的移动进行更加鲁棒的预测,即使在目标外观遮挡、光照变化、快速运动等复杂情况下也能保持较好的跟踪性能。追踪器会为每个被检测到的行人分配唯一的 ID,并在下一帧中根据检测置信度和跟踪策略来判断是否保持或删除该 ID。经以上方法,所获取的行人位置信息向量 $\boldsymbol{S}_k$ 表示为

$$\boldsymbol{S}_k=(x^i,y^i,w^i,h^i)^{\mathrm{T}} \tag{3}$$

式(3)中,i 表示 k 时刻中所被跟踪的行人 ID。(后文所有参数下标 k 表示时刻,上标 i 表示行人 ID,由于篇幅关系不再赘述。)

2.2　身体信息获取

使用 YOLO-Pose[15] 拟合行人骨架,得到行人头部及肢体共 17 个关键点,$\boldsymbol{P}_k$ 记录行人在时间序列内关键点的动态变化,即:

$$\boldsymbol{P}_k=(a_1^i,b_1^i,\cdots,a_u^i,b_u^i,\cdots,a_{17}^i,b_{17}^i)^{\mathrm{T}} \tag{4}$$

式(4)中,(a_u^i,b_u^i) 为行人的第 u 关键点在 k 时刻被拟合的横纵坐标。

考虑姿态估计方法对交通环境中的小目标拟合不佳,采用数据预处理与后处理方法增强数据。

(1) 预处理。预处理的目的是通过对目标图像进行尺度放大以增加图像中行人的可辨识性。设帧宽为 W_{frame},将每帧被检测到的行人在原始图片中按照边界框的 w_k^i,h_k^i 进行裁剪,对被裁剪下来的小目标($w_k^i/W_{\mathrm{frame}}\leqslant$

0.023)和中目标(0.023<w_k^i/W_{frame}≤0.035)采用等比例放大的方法,放大的策略为小目标(×3),中目标(×2)。将每个处理后的目标边界框($\bar{w}_k^i$,$\bar{h}_k^i$)基于 Maximal Rectangles Bottom-Left[16]算法进行处理,使其以最大的空间利用率放入宽高(W_k,H_k)为式(5)所示的矩形空间中。这些边界框在矩形空间中占用的最大矩形尺寸为(W_k,H_k),并将最大占用的矩形图片作为 YOLO-Pose 的输入,输出的行人骨架则通过对应的 ID 信息缩小还原在原始图片中。

$$(W_k,H_k)=\begin{cases}(\bar{w}_k^1,\bar{h}_k^1) & ,n=1\\ \left(\max\left(\left\lceil\dfrac{\sum_{i=1}^{n}\bar{w}_k^i}{2}\right\rceil,\bar{w}_k^1\cdots\bar{w}_k^n\right),\sum_{i=1}^{n}\bar{h}_k^i\right) & ,n\geqslant 2\end{cases} \tag{5}$$

(2)后处理。因小目标裁剪放大造成目标模糊,从而导致行人时序骨架的拟合噪声较大。为消除该影响,采用 RTS(Rauch-Tung-Striebel)平滑算法[17]和抗差 Kalman 滤波[18]的数据后处理方法。以行人在连续帧区间内关节点的坐标以及其对应的瞬时变化速度作为全局待估状态向量 $\boldsymbol{X}_k$,表示为

$$\boldsymbol{X}_k=(\boldsymbol{P}_x^{\,i},\boldsymbol{P}_y^{\,i},\boldsymbol{V}_x^{\,i},\boldsymbol{V}_y^{\,i})^{\mathrm{T}} \tag{6}$$

式(6)中,$\boldsymbol{P}_x$ 与 $\boldsymbol{P}_y$ 分别表示行人骨架的横坐标和纵坐标,$\boldsymbol{V}_x$ 与 $\boldsymbol{V}_y$ 表示行人的骨架点分别沿着横和纵坐标的瞬时速度。

反向过程利用历史正向滤波的部分数据再次滤波,对时序骨架进行输出校正,以获取更精确的骨架节点估计值。经 Kalman 滤波对数据做抗差处理后,初始 $k=t$,方向为负,利用反向滤波再对该段数据做 RTS 固定区间最优平滑,时间区间为[t,1],递推公式如下:

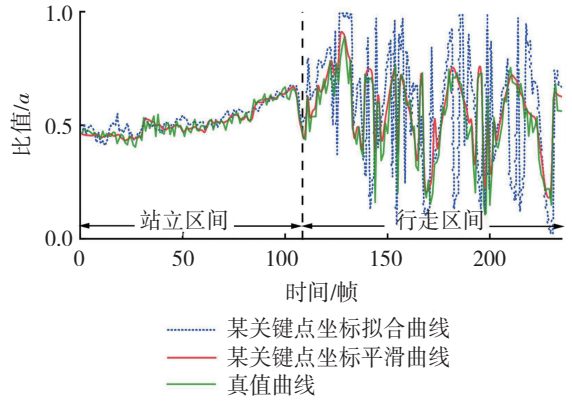
$$\boldsymbol{K}_{k-}=\boldsymbol{Q}_k^{+}\boldsymbol{\Lambda}_k^{\mathrm{T}}(\boldsymbol{Q}_{k+1}^{-})^{-1} \tag{7}$$

$$\hat{\boldsymbol{X}}_k=\hat{\boldsymbol{X}}_k^{+}+\boldsymbol{K}_{k-}(\hat{\boldsymbol{X}}_{k+1}-\hat{\boldsymbol{X}}_{k+1}^{-}) \tag{8}$$

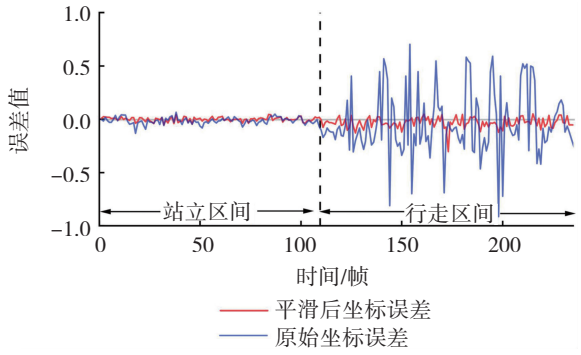
$$\boldsymbol{Q}_k=\boldsymbol{Q}_k^{+}+\boldsymbol{K}_{k-}(\boldsymbol{Q}_{k+1}-\boldsymbol{Q}_{k+1}^{-})\boldsymbol{K}_{k-}^{\mathrm{T}} \tag{9}$$

式(7)—式(9)中,$\boldsymbol{\Lambda}_k$ 表示状态转移矩阵;$\boldsymbol{K}_{k-}$ 表示反向滤波增益矩阵;$\hat{X}_k$ 与 $\boldsymbol{Q}_k$ 分别表示在反向滤波过程中待估状态向量及其协方差矩阵;$\hat{X}_{k+1}^{-}$ 与 $\hat{X}_{k+1}^{+}$ 分别表示在正向滤波过程中先验与后验估计状态向量;$\boldsymbol{Q}_k^{-}$ 与 $\boldsymbol{Q}_k^{+}$ 分别表示在正向滤波过程中先验与后验估计状态的协方差矩阵。

图 2 展示了某段视频序列中行人左脚关键点坐标的平滑结果。其中,图 2(a)纵坐标参数 a 为当前帧行人关键点横坐标与检测边界框宽的比值,该参数可以在人车相对移动的场景中达到归一化目的。从图 2(b)的误差对比看出,经过滤波平滑校正的时序骨架可以有效地减少姿态拟合的误差。



(a) 骨架关键点(左脚)横坐标的平滑效果



(b) 平滑前后的误差对比

图 2 行人关键点坐标的平滑效果

Fig. 2 Smoothing effect of pedestrian keypoint coordinates

2.3 动作信息获取

采用基于骨架的动作识别算法 ST-GCN++[19]识别行人动作。该算法以行人在时间序列 t(根据文献[9],本文将 t 设为 20)中的动态骨架作为输入;在进行批量标准化后,将数据分为空间边和时间边,并分别经过 GCN 图卷积和 TCN 时间卷积的处理;最后,特征图通过池化层和全连接层输出识别结果。

鉴于行人动作表现的高度灵活性,受 FineGym 数据集[20]的启发,动作识别器将行人时序骨架的动作分为 st(站立)、st-at(站立-注意力)、wa(行走)、wa-at(行走-注意力)四种基础步态行为。其中 at(注意力)态指的是行人在扭头、转身或挥手致意的过程中能够注意到主车的状态。则动作信息向量 $\boldsymbol{A}_k$ 表示为

$$\boldsymbol{A}_k=(\eta_{\text{st}}^i,\eta_{\text{st-at}}^i,\eta_{\text{wa}}^i,\eta_{\text{wa-at}}^i)^{\mathrm{T}} \tag{10}$$

式(10)中,η_{st}、$\eta_{\text{st-at}}$、η_{wa}、$\eta_{\text{wa-at}}$ 分别对应站立、站立-注意力、行走、行走-注意力的概率。

2.4 环境信息获取

使用 PIDNet[21]提取周边环境信息。所获取的环

境信息向量 $\boldsymbol{E}_k$ 表示为

$$\boldsymbol{E}_k=(r(X,Y))^{\mathrm{T}} \tag{11}$$

式(11)中,$r(\ )$为环境像素统计函数,返回定义域内所统计目标像素的总数,其定义域为$\{(X,Y)\mid 1\leqslant X\leqslant W_{\mathrm{frame}},1\leqslant Y\leqslant H_{\mathrm{frame}}\}$。文中目标像素为车辆可行驶路面。

3　基于多特征融合的穿越行为预测

为了消除多人场景中的预测噪声,本文提出了先条件后预测(Predict after Condition,PAC)的两阶段方法。PAC 方法分为先决条件判断与特征融合预测两部分,如图 3 所示。

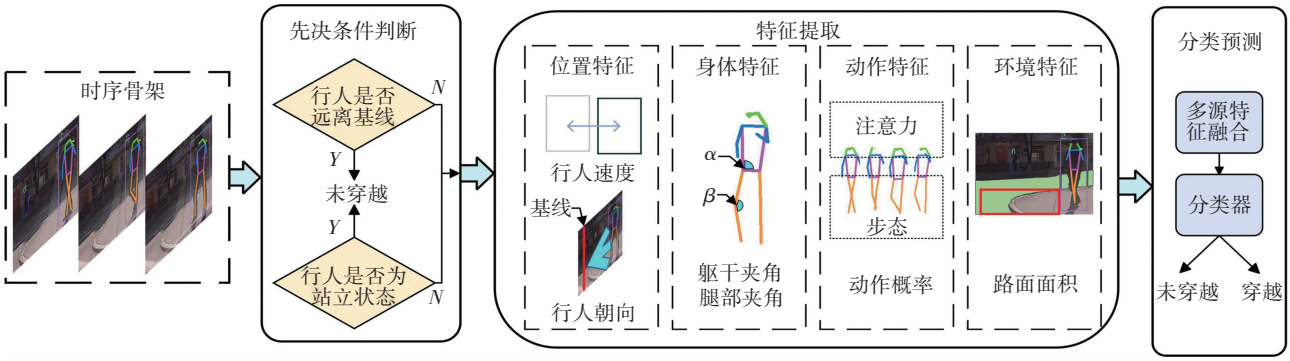


图 3　先条件后预测的两阶段方法

Fig. 3　Two-stage method of PAC

在视频横轴 $W_{\mathrm{frame}}/2$ 处做一条垂线作为基线,通过在车载视频中引入“基线”建立人车之间的关系,利用先决条件(行人的步态信息以及靠近基线的趋势)剔除不具备穿越条件的行人,并将剩下可能穿越的行人交给分类器进行预测。将历史信息长度设为 m,各信息特征的提取如下。

3.1　位置特征提取

(1) 行人速度。以视频中行人的横向轨迹进行分析,取当前帧的行人边界框中心横坐标 $\tilde{x}_k$,减去历史中心横坐标 $\tilde{x}_{k-m}$,取平均后,除以区间平均边界框宽进行标准化,得到特征值由式(12)表示:

$$\Delta v_k=\frac{m}{m-1}\frac{\tilde{x}_k-\tilde{x}_{k-m}}{\sum_{j=0}^{m}w_{k-j}} \tag{12}$$

(2) 行人朝向。设特征值 q_k 用于记录行人朝向情况。若行人速度为正,则将当前帧 $\tilde{x}_k<W_{\mathrm{frame}}/2$ 的目标记为 1(左侧目标正在向基线靠近),并参与预测;若行人速度为负,则将当前帧 $\tilde{x}_k>W_{\mathrm{frame}}/2$ 的目标记为 0(右侧目标正在向基线靠近),且参与预测;其余情况不予标记预测。

综上,位置特征向量表示为 $\hat{\boldsymbol{S}}_k=(\Delta v^i,q^i)^{\mathrm{T}}$。

3.2　身体特征提取

(1) 躯干夹角。把行人身体边组成的四边形作为躯干部分,将躯干的四个内角在连续 m 帧中的变化量作为躯干变化特征,四角按照从小到大的顺序编号为右肩、右胯、左胯和左肩。设当前帧躯干的某个夹角为 α_k,则连续帧中的变化量可由式(13)表示:

$$\Delta\alpha_k=\sum_{j=0}^{m-1}|\alpha_{k-j}-\alpha_{k-j-1}| \tag{13}$$

(2) 腿部夹角。腿部特征以行人大小腿之间的夹角在连续 m 帧中的变化量表示。设当前帧腿部夹角为 β_k,则连续帧中变化量可由式(14)表示:

$$\Delta\beta_k=\sum_{j=0}^{m-1}|\beta_{k-j}-\beta_{k-j-1}| \tag{14}$$

综上,身体特征向量表示为 $\hat{\boldsymbol{P}}_k=(\Delta\alpha^i,\Delta\beta^i)^{\mathrm{T}}$。

3.3　动作特征提取

本文将行人在穿越过程中是否具有关注主车的行为,划分为谨慎型和冒进型两类。如图 4 所示,根据统计样本中行人的行走-注意力概率在穿越时明显下降的趋势,取其概率作为动作特征,即:$\hat{\boldsymbol{A}}_k=(\eta^i_{\mathrm{wa-at}})^{\mathrm{T}}$。

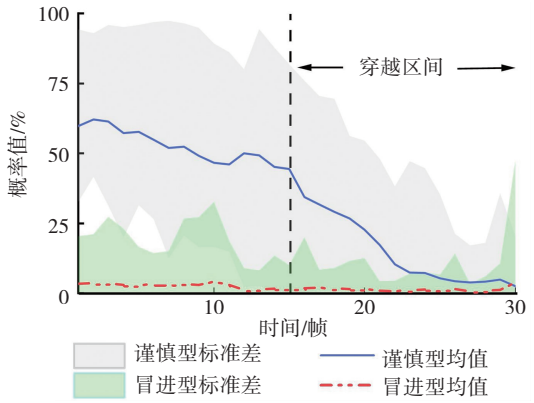


图 4　行人在穿越过程中行走-注意力概率的变化情况

Fig. 4　The variation of pedestrian actions in terms of walking and attention probability during the crossing process

3.4　环境特征提取

路面面积取行人脚下局部的矩形区域,计算其连

续 m 帧内包含道路的平均占比,即:

$$\Delta e_k = \frac{1}{m}\sum_{j=0}^{m-1}\frac{r(X,Y)}{w_{k-j}\times h_{k-j}} \tag{15}$$

式(15)中,$r(\ )$ 函数的定义域根据当前行人朝向确定。当 $q_k = 1$ 时,$\tilde{x}_k + w_k/2 \leqslant X \leqslant \tilde{x}_k + (w_k+2h_k)/2$ 且 $\tilde{y}_k + (h_k-w_k)/2 \leqslant Y \leqslant \tilde{y}_k+(h_k+w_k)/2$;当 $q_k=0$,$\tilde{x}_k-w_k/2 \leqslant X \leqslant \tilde{x}_k-(w_k+2h_k)/2$ 且 $\tilde{y}_k+(h_k-w_k)/2 \leqslant Y \leqslant \tilde{y}_k+(h_k+w_k)/2$。

则环境特征向量 $\hat{\boldsymbol{E}}_k = (\Delta e^i)^{\mathrm{T}}$。

4 仿真实验与结果分析

4.1 实验数据集获取

采用自动驾驶联合注意力公开数据集 JAAD(Joint Attention in Autonomous Driving Public Dataset)[7] 用于实验的训练和分析。该数据集包含 346 个高分辨率的视频剪辑,涵盖了多种人车交汇场景,例如:十字路口、城市或乡间道路等。弱势道路使用者可分为存在交汇意图的行人和无交汇意图的行人两种,考虑到前者多存在于无遮挡或少遮挡(25%至 75%可见)情况,且拥有更多的步态、注意力、穿越等信息。因此在训练环节中,使用前者提取时序骨架并训练动作识别器,同时使用两者训练检测器和分类器。

数据集划分方案如表 1 所示,其中检测器的数据样本为抽帧图片,动作识别器与分类器分别为时序骨架数据和归一化的多特征数据。

表 1 训练集与测试集的划分

Table 1 Division of training set and test set

方 法	识别类别	训练集	测试集
检测器	行人	51 072	21 000
动作识别器	站立	98	20
	站立-注意力	106	20
	行走	117	20
	行走-注意力	92	20
分类器	未穿越	3 507	1 503
	穿越	3 507	1 503

4.2 评价指标

实验的评价指标采用准确率(Accuracy,σ_{Acc}),精确率(Precision, P),召回率(Recall,R),F1 分数(F_1),平均精度(Average Precision, σ_{AP})综合评估模型性能,计算公式分别为

$$\sigma_{\mathrm{Acc}} = \frac{T_{\mathrm{P}}+T_{\mathrm{N}}}{T_{\mathrm{P}}+T_{\mathrm{N}}+F_{\mathrm{P}}+F_{\mathrm{N}}} \tag{16}$$

$$P = \frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{17}$$

$$R = \frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \tag{18}$$

$$F_1 = \frac{2\times P\times R}{P+R} \tag{19}$$

式(16)—式(19)中,T_{P} 和 T_{N} 分别表示正确预测的未穿越样本和穿越样本;F_{P} 和 F_{N} 分别表示错误预测的未穿越样本和穿越样本。

$$\sigma_{\mathrm{AP}} = \int_0^1 p(r)\,\mathrm{d}r \tag{20}$$

式(20)中,p 代表 Precision ; r 代表 Recall。

4.3 推理细节

模型的推理规则如图 5 所示,本文提出的模型通过利用历史信息进行推理,并随着时间迭代更新后续的预测。整个预测过程分为三个阶段:在第一阶段,由于尚未获取足够的历史信息,模型不进行任何预测输出;在第二阶段,当时序骨架长度小于动作识别所需的动态骨架长度时,动作特征将无法提取,模型将在无动作特征的情况下进行预测;而在第三阶段,所有特征均被正常提取,模型进入正常的预测状态。

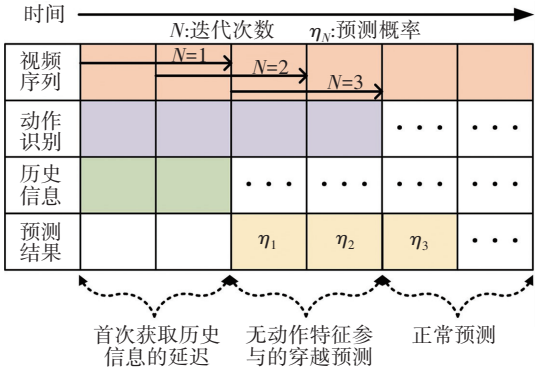


图 5 模型的推理规则

Fig. 5 Inference rules of the model

4.4 实验结果分析

图 6 展示了在正负样本均衡的测试样本集 A、B、C、D 中,模型基于不同长度历史信息进行特征提取时的准确率表现。其中样本集 A、B、C、D 分别选取了行人在穿越前后 3 s、2 s、1 s、0.5 s 的视频序列。观察结果可见,在样本集 A 中,随着历史序列长度的增加,预测的准确率逐渐提高。然而,在样本集 B、C、D 中,准确率会随着序列长度的增长达到一个峰值后开始逐渐降低,这三个样本集的最优序列长度分别为 5 帧、5 帧和 3 帧。这一现象的原因在于模型依赖历史信息进行预测。当行人的检测时间较短时,增加历史序列长度会导致首次获取历史信息的延迟和预测误差所占比重较

大;相比之下,当行人的检测时长较长时,这些延迟和误差的影响则显得较为微小。鉴于实际情况和预测精度的综合考量,本研究选择采用长度为 5 的历史信息来训练和测试模型。

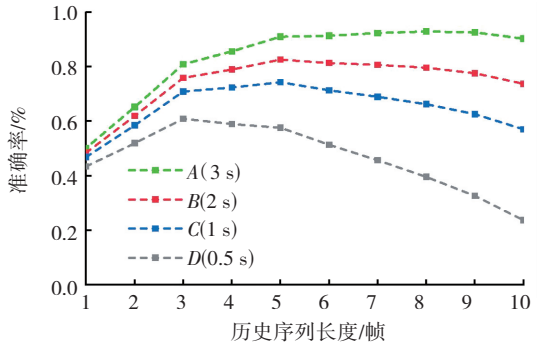


图 6　选取不同长度历史信息对精度的影响

Fig. 6　The impact of selecting different lengths of historical information on accuracy

表 2 展示了在不同的分类器上,未经先决条件筛选直接参与预测的“单阶段”方法与“两阶段”方法的测试精度对比。

表 2　单阶段与两阶段方法在不同分类器上的预测结果

Table 2　Prediction results of one-stage and two-stage methods on different classifiers　/%

分类器	训练形态	精准率	平均精度
Perceptron	单阶段	56.77	53.13
	两阶段	97.76	88.63
SVM	单阶段	88.39	82.30
	两阶段	90.20	84.13
Random Forest	单阶段	85.38	81.15
	两阶段	92.96	88.98
XGBoost (Ours)	单阶段	84.88	80.55
	两阶段	93.29	89.31

从表 2 中可以看出,无论是在线性分类器(Perceptron、SVM)还是非线性分类器(Random Forest、XGBoost)上,采用两阶段方法的平均精度都较单阶段有所提升。其中,在 Perceptron、RandomForest、XGBoost 分类器上,平均精度的提升尤为明显,分别提高了 35.50%、7.83%、8.76%;在 XGBoost 分类器上,平均精度达到最高值 89.31%。

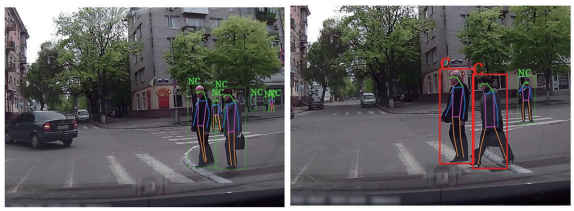
表 3 所示为模型在 JAAD 数据集上的定量实验结果,通过对比其他模型可知,本文的模型相较于 ResNet[22] 方法的平均精度提高了 2.19%;在与最新模型 TrEP[23] 预测准确率相近的情况下,本文的模型表现出更强的稳健性。通过消融对比,经过 RTS 平滑处理的时序骨架在 C/NC 预测中相较于未处理的情况,平均精度提高了 1.43%。

表 3　各模型在 JAAD 数据集上的对比

Table 3　Comparison of models on the JAAD dataset　/%

模　型	准确率	F_1 分数	平均精度
2s-AGCN[11]	86.72	82.95	71.32
ResNet[22]	90.31	89.41	87.12
PCPA[8]	87.21	79.42	85.38
TrEP[23]	91.81	70.12	81.23
Ours(未平滑)	91.31	90.01	87.88
Ours(平滑)	91.55	91.78	89.31

图 7 为模型在驾驶员的不同决策下对行人穿越意图的预测结果,图中 C 表示穿越,NC 表示未穿越。从图中的两个案例可以观察到,本文模型所采用的 PAC 方法仅对那些靠近基线并处于行走状态的行人进行了后续预测。对于位置远离基线或处于站立状态的行人,模型会直接将其分类为 NC。这一决策基于这样的事实:在车辆的前视视角下,如果行人未来的横向轨迹无法与主车交汇,那么将其判定为 NC 将更具有预测意义。



(a) 车辆礼让

(b) 车辆未礼让

图 7　模型在驾驶员不同决策下的预测结果

Fig. 7　Model predictions under different driver decisions

然而,PAC 方法在处理由车辆转向运动所引起的视觉突变场景时存在一定的局限性。这种视觉上的横向突变可能导致人车间通过基线建立的关联与原始预测模型产生冲突。为了解决这一问题,未来的研究可以考虑在模型中引入自适应基线机制。

同时对特征的重要度进行分析。如图 8 所示,通过统计训练过程中特征被调用的次数,选取出排名前三的特征:行人速度、动作概率、路面面积。

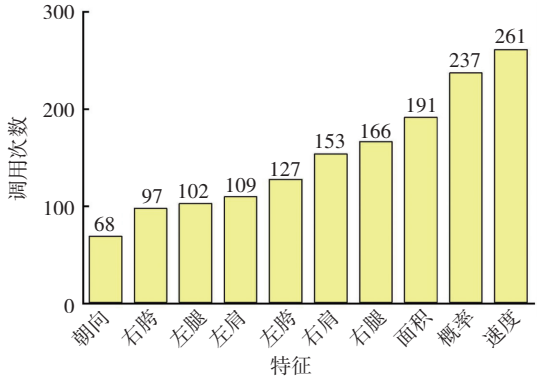


图 8 特征在训练过程中被调用的次数

Fig. 8 Feature utilization frequency during the training process

将以上三个特征与剩余特征以不同的组合方式进行测试,如图 9 所示。

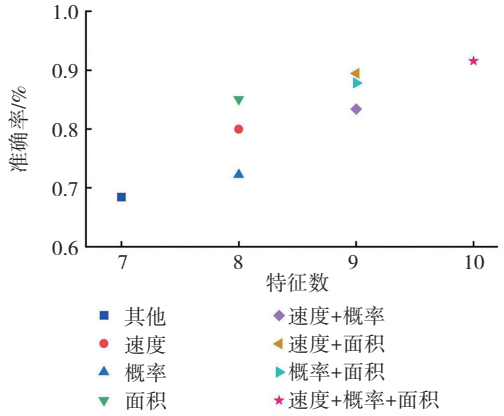


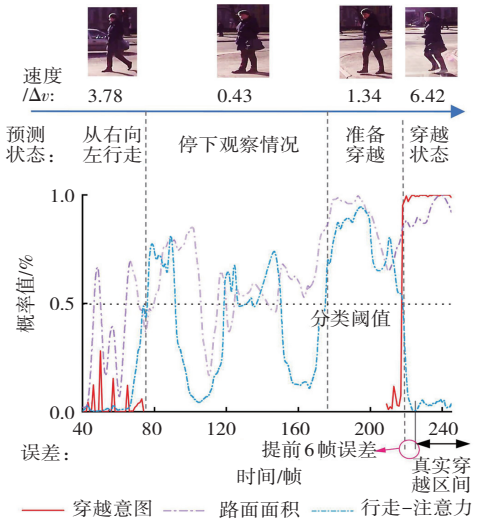
图 9 不同重要特征搭配对预测准确率的影响

Fig. 9 The impact of different combinations of important features on prediction accuracy

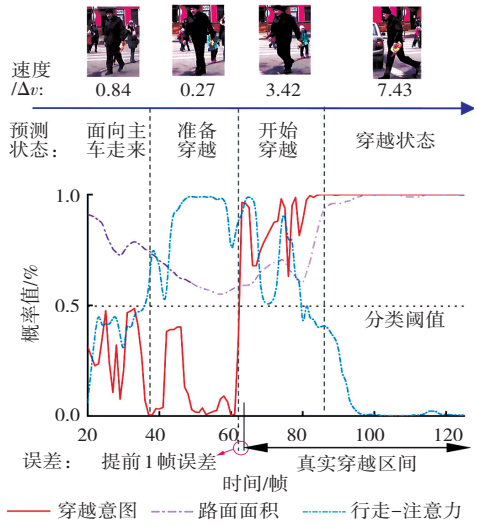
经分析,在三者未参与的情况下,预测的准确率仅为 68.43%。当加入路面面积这一特征时,预测准确率提升至 85.06%。如果同时考虑路面面积和行人速度,准确率进一步提高到 89.45%。而当这 3 个特征共同作用时,预测准确率达到了最高值,即 91.55%。因此,可以得出结论,路面面积、行人速度和行走-注意力概率是影响行人穿越行为的 3 个重要特征。

为了测试模型对不同类型行人的泛化能力,考虑行人穿越前动作可分为站立和行走两种状态,结合谨慎型和冒进型的分类,将行人细分为以下四类:谨慎型-站立态、谨慎型-行走态、冒进型-站立态和冒进型-行走态。

图 10 和图 11 展示了本文所提模型在 JAAD 数据集上结合重要特征的分析结果,以 0.5 作为分类阈值判断穿越(C)/未穿越(NC)意图。



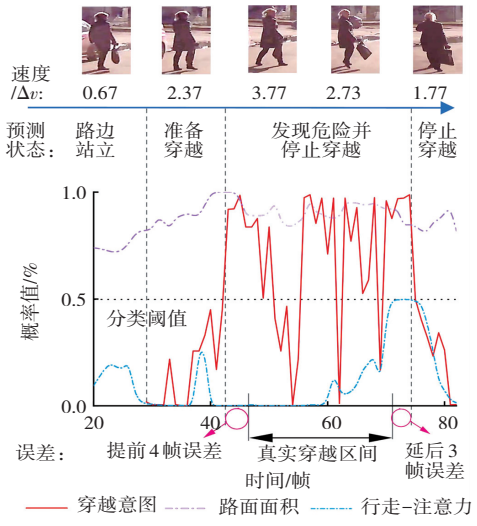
(a) 谨慎型-站立态行人



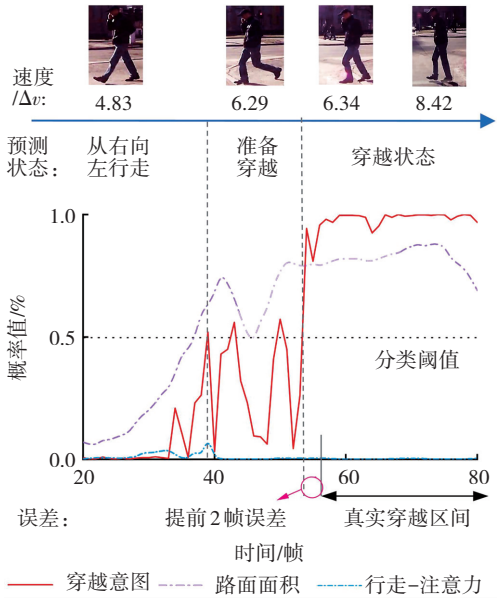
(b) 谨慎型-行走态行人

图 10 谨慎型行人的穿越行为预测及误差

Fig. 10 Prediction and error of crossing behavior of cautious pedestrian



(a) 冒进型-站立态行人



(b) 冒进型-行走态行人

图 11　冒进型行人的穿越行为预测及误差

Fig. 11　Prediction and error of crossing behavior of aggressive pedestrian

分析图 10 和图 11 可知,在满足交汇条件的前提下,脚下路面面积超过 65%且速度超过 3 的行人穿越意图较强烈。在图 10(a)和图 11(a)中,当行人的预测状态为“停下观察情况”或“路边站立”时,模型因先决条件判断直接在预测结果上将行人不可能穿越的“站立”态判定为 NC,从而消除了该阶段的预测噪声。在图 11(a)“发现危险并停止穿越”的预测状态中,行人实际处于穿越又因感知危险而减速的状态,因此预测的概率呈现出震荡的趋势;在误差方面,预测结果与标注的真实值相比存在略微的“提前误差”和“延后误差”。本研究认为这些略微的误差是模型对穿越行为敏感的表现,因为与预测延迟相比,预测提前更可被接受。

综上,基线机制的引入使得在车载前视视角环境下,可以通过分析行人在观测场景中的横向轨迹来建立人车之间的关系,并由此衍生出 PAC 方法使得预测更加具有针对性和实用性。通过结合行人的位置、身体、动作以及局部环境特征进行多特征建模,进一步提升了预测的准确性。

5　结　论

采用多种方法提取交通环境中的多源信息特征,并使用分类器融合这些特征来预测行人的穿越概率。针对复杂环境下造成的行人骨架信息丢失问题,采用了不同尺度放大的预处理和数据滤波平滑的后处理措施,通过对图像进行尺度放大,可以增加图像中行人的可辨识性,同时,数据滤波平滑可以减少噪声对特征提取的影响,进一步提高预测模型的性能。针对车辆前视视角下的穿越行为,提出了 PAC 两阶段方法。

基于 JAAD 数据集下的实验结果表明:经 RTS 平滑处理的模型相较未处理的情况平均精度提升了 1.43%;采用 PAC 方法较传统单阶段方法平均精度提高了 8.76%,达 89.31%。在特征重要性方面,当加入路面面积这一特征时,预测准确率由 68.43%提升至 85.06%,由此可知,考虑行人位置与路面轮廓、路缘带之间的关系是研究行人穿越行为的重要因素。

参考文献(References):

[1] 杨永斌, 李笑扬. 基于大数据技术的智能交通管理与应用研究[J]. 重庆工商大学学报(自然科学版), 2019, 36(2): 73-79.
YANG Yong-bin, LI Xiao-yang. Research on intelligent traffic management and application based on big data technology[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2019, 36(2): 73-79.

[2] 张亚丽. 世界卫生组织发布《2018 年全球道路安全现状报告》[J]. 中华灾害救援医学, 2019, 5(2): 100.
ZHANG Ya-li. The World Health Organization released the report on global road safety status in 2018[J]. Chinese Journal of Disaster Medicine, 2019, 5(2): 100.

[3] ALAHI A, GOEL K, RAMANATHAN V, et al. Social LSTM: human trajectory prediction in crowded spaces[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 961-971.

[4] HUG R, BECKER S, HÜBNER W, et al. Particle-based pedestrian path prediction using LSTM-MDL models [C]// Proceedings of the 21st International Conference on Intelligent Transportation Systems. Piscataway: IEEE Press, 2018: 2684-2691.

[5] 李文礼, 张祎楠, 王梦昕. 基于视野域机制的行人轨迹预测[J]. 计算机应用研究, 2023, 40(1): 80-85.
LI Wen-li, ZHANG Yin-an, WANG Meng-xin. Pedestrian trajectory prediction based on field of view mechanism [J]. Application Research of Computers, 2023, 40(1): 80-85.

[6] 陈龙, 杨晨, 蔡英凤, 等. 基于多模态特征融合的行人穿越意图预测方法[J]. 汽车工程, 2023, 45(10): 1779-1790.

CHEN Long, YANG Chen, CAI Ying-feng, et al. Pedestrian crossing intention prediction method based on multimodal feature fusion[J]. Automotive Engineering, 2023, 45(10): 1779-1790.

[7] RASOULI A, KOTSERUBA I, TSOTSOS J K. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE Press, 2017: 206-213.

[8] KOTSERUBA I, RASOULI A, TSOTSOS J K. Benchmark for evaluating pedestrian action prediction[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2021: 1257-1267.

[9] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proc AAAI Conf Artificial Intelligence. Piscataway: AAAI Press, 2018: 7444-7452.

[10] CHEN T, TIAN R, DING Z. Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE Press, 2021: 3096-3102.

[11] 胡远志, 蒋涛, 刘西, 等. 基于双流自适应图卷积神经网络的行人过街意图识别[J]. 汽车安全与节能学报, 2022, 13(2): 325-332.
HU Yuan-zhi, JIANG Tao, LIU Xi, et al. Pedestrian-crossing intention-recognition based on dual-stream adaptive graph-convolutional neural-network[J]. Journal of Automotive Safety and Energy, 2022, 13(2): 325-332.

[12] 李习习, 强俊, 刘无纪, 等. 基于双主干网络的雾天交通目标检测方法研究[J]. 重庆工商大学学报(自然科学版), 2023, 40(4): 25-34.
LI Xi-xi, QIANG Jun, LIU Wu-ji, et al. Research on traffic object detection method in fog based on dual backbone network[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(4): 25-34.

[13] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 7464-7475.

[14] ZHANG Y, SUN P, JIANG Y, et al. ByteTrack: Multi-object tracking by associating every detection box[C]//Avidan S, Brostow G, Cissé M, et al. European Conference on Computer Vision. Cham: Springer, 2022: 1-21.

[15] MAJI D, NAGORI S, MATHEW M, et al. YOLO-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2022: 2636-2645.

[16] CHAZELLE. The bottomn-left Bin-packing heuristic: An efficient implementation[J]. IEEE Transactions on Computers, 1983, C-32(8): 697-707.

[17] SÄRKKÄ S. Unscented rauch: tung: striebel smoother[J]. IEEE Transactions on Automatic Control, 2008, 53(3): 845-849.

[18] KALMAN R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960, 82(1): 35-45.

[19] DUAN H, WANG J, CHEN K, et al. PYSKL: Towards good practices for skeleton action recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 7351-7354.

[20] SHAO D, ZHAO Y, DAI B, et al. FineGym: A hierarchical video dataset for fine-grained action understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2613-2622.

[21] XU J, XIONG Z, BHATTACHARYYA S P. PIDNet: A real-time semantic segmentation network inspired by PID controllers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 19529-19539.

[22] RAZALI H, MORDAN T, ALAHI A. Pedestrian intention prediction: A convolutional bottom-up multi-task approach[J]. Transportation Research Part C: Emerging Technologies, 2021, 130: 103259.

[23] ZHANG Z, TIAN R, DING Z. TrEP: Transformer-based evidential prediction for pedestrian intention with uncertainty[C]//Proc AAAI Conf Artificial Inte-lligence. Piscataway: AAAI Press, 2023: 3534-3542.

责任编辑:陈 芳