中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于多信息融合的驾驶视角下行人轨迹预测

桑海峰, 刘泉恺, 王金玉, 陈旺兴

引用本文:
桑海峰,刘泉恺,王金玉,陈旺兴. 基于多信息融合的驾驶视角下行人轨迹预测[J]. 控制与决策, 2024, 39(7): 2354−2362.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.2229

基于多信息融合的驾驶视角下行人轨迹预测

桑海峰†, 刘泉恺, 王金玉, 陈旺兴

(沈阳工业大学 信息科学与工程学院，沈阳 110870)

摘 要: 行人轨迹预测是实现在城市内完全自动驾驶的重要支撑,并且广泛应用于机器人路径规划、自主巡航等领域. 驾驶视角下交通场景复杂多变、行人未来位置不确定性大,只考虑观测轨迹信息预测行人轨迹会有较大位移误差. 针对这个问题,提出一种多信息融合网络 (multi-information fusion network, MIFNet) 来预测驾驶视角下未来行人轨迹的多种可能. MIFNet 在观测轨迹信息的基础上引入姿态信息和光流信息,分别采用骨架序列重组和划分局部光流的方法避免遮挡造成的信息失真. 为了更有效地融合这些信息,提出一种基于信息评价的跨信息融合注意力机制,综合考虑了预测过程中不同信息间的重要程度和同一信息间不同特征的重要程度. MIFNet 在 PIE 数据集上预测 1.5 s 的平均位移误差取得了最佳成绩,在 JAAD 数据集 1.5 s 的长时轨迹预测任务中预测误差最小,并且模型参数量、推理时间较最新模型大幅度下降.

关键词: 行人轨迹；轨迹预测；驾驶视角；多信息融合；注意力机制；信息评价

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.2229

引用格式: 桑海峰, 刘泉恺, 王金玉, 等. 基于多信息融合的驾驶视角下行人轨迹预测 [J]. 控制与决策, 2024, 39(7): 2354-2362.

Pedestrian trajectory prediction from driving perspective based on multi-information fusion

SANG Hai-feng†, LIU Quan-kai, WANG Jin-yu, CHEN Wang-xing

(School of Information Science and Engineering，Shenyang University of Technology，Shenyang 110870，China)

Abstract: Pedestrian trajectory prediction is an important support for fully automatic driving in the city, and is widely used in robot path planning, autonomous cruise and other fields. The traffic scene from the driving perspective is complex and changeable, and the future position of pedestrians is uncertain. In this paper, a multi-information fusion network (MIFNet) is proposed to predict multiple possibilities of future pedestrian trajectories. Pedestrian posture information and optical flow information are used in the MIFNet on the basis of observed trajectory information, and the ways of reconstructing skeleton sequences and dividing local optical flow are used to avoid information distortion caused by pedestrian occlusion. In order to fuse these information more effectively, this paper proposes a cross-information fusion attention mechanism based on information evaluation. The importance of different information in the prediction process and the importance of different features between the same information are comprehensively considered. The MIFNet achieves the best results in predicting the average displacement error of 1.5 seconds on the PIE dataset, and the long-term trajectory prediction task of 1.5 seconds on the JAAD dataset. The prediction error is the smallest, and the number of model parameters and inference time are greatly reduced compared with the latest model.

Keywords: pedestrian trajectory；trajectory prediction；driving perspective；multi-information fusion；attention mechanism；information evaluation

0 引 言

近年来,随着新能源汽车产业的蓬勃发展,大量的汽车开始搭载自动驾驶技术. 但是已经量产的自动驾驶车辆使用场景大多是高速公路、城市高架桥等道路构成元素较为简单,路上行人较少的交通场景,在行人密集、道路环境复杂的城市街道还处于测试环节. 为了得到更安全稳健的自动驾驶决策,预测道路行人的运动轨迹和意图是至关重要的一环. 行人轨迹预测是避免碰撞的关键,通过预测道路上行人的轨迹,智能汽车可以在交通中规划安全且具有社会

收稿日期: 2022-12-31；**录用日期:** 2023-05-10.

基金项目: 国家自然科学基金项目 (62173078)；辽宁省自然科学基金项目 (2022-MS-268).

责任编委: 侯忠生.

†通讯作者. E-mail: sanghaif@163.com.

意识的路径,并对异常运动发出警报,有效地避免碰撞发生[1]. 驾驶视角下利用车载相机采集行人视频,在不借助激光雷达等传感器的条件下可以实现低成本的行人轨迹预测. 与鸟瞰视角不同,在驾驶视角下可以近距离捕捉行人姿态,从而利用行人姿势来预测他们的未来位置. 同时,驾驶视角下行人的大小不可忽视,将其简化为空间中的一个点只考虑行人历史轨迹信息会有较大预测误差. 车载相机跟随汽车一起运动,在行人轨迹预测中其自我运动往往不能忽略[2]. 此外,行人和周围的行人、运动的汽车、固定的建筑物等之间的相互遮挡不可避免,这对观测行人姿态和预测轨迹带来极大的挑战. 针对以上问题,本文的主要创新点如下:

1) 提出一种多信息融合的行人轨迹预测模型MIFNet,较为全面地考虑行人的姿态信息、光流信息以及观测轨迹信息,并利用所提出的基于信息评价的跨信息融合网络有效地融合这些信息.

2) 为了解决行人姿态信息识别中出现的遮挡问题,提出一种基于置信度信息的骨骼关键帧序列提取方法,以减少行人骨架序列中无效帧的影响.

3) 利用行人光流变化表示相机运动对轨迹预测的影响,为了减少光流估计过程中外界遮挡的影响,提出一种基于CNN-RNN时空结合的局部光流编码网络.

1　相关工作

行人轨迹预测按照视角不同划分为鸟瞰视角和驾驶视角. 在鸟瞰视角下,建模较为简单,且研究较早,相关研究工作较多,已经有比较成熟的体系和较低的预测误差. 驾驶视角场景下的行人轨迹预测研究起步晚,考虑因素多,建模难度较大. 行人轨迹预测的发展方向可以分为: 从确定性模型(单条预测轨迹)到生成概率模型(多模态轨迹输出); 从单信息源模型(观测轨迹信息)到多信息源模型. 本节主要根据后者的相关研究顺序进行阐述.

1.1　单信息源行人轨迹预测

早期的轨迹预测工作通常假设未来是确定的,根据过去的观测结果,采用循环神经网络模型(recurrent neural network, RNN)进行轨迹预测,每个行人只能预测一条轨迹. 然而,行人的移动具有高度的随机性,因此可能存在多种可能的和不同的未来行为[3]. 最近的研究表明,采用生成对抗网络、条件变分自编码器、扩散模型等生成概率模型预测多条潜在未来轨迹的分布而不是单个最佳轨迹可以更准确地模拟行人的未来运动[4-6]. 针对循环神经网络对于时间推移(> 560 ms)带来的模型恶化问题,文献[7]提出了基于远期目标估计的轨迹预测模型; 文献[8]提出一种考虑出行方式的基于马尔可夫模型的轨迹预测方法,按照出行方式对轨迹分类; 文献[9]根据人类运动可能受周围环境影响从而改变目标位置的特性,提出了逐步目标估计模型,将远期目标分为多个短期目标进行轨迹预测. 仅通过历史轨迹信息较难在高度动态场景组成的城市交通环境中准确预测行人未来的轨迹与意图,因此本文考虑行人姿态、光流等多种信息来预测行人轨迹.

1.2　多信息源行人轨迹预测

基于单信息的预测模型在交通场景复杂、预测长时轨迹的过程中较难准确估计行人的短期意图和运动轨迹,因此在观测轨迹信息的基础上引入其他信息成为近些年的研究热点. 文献[2]中提取行人的骨架信息,用来表征行人的姿态,从而预测行人未来位置,但是忽视了提取骨架信息过程中因骨骼关键点残缺产生无效序列帧的问题; 文献[10]提出了一种结合行人意图和车速预测的行人轨迹预测方法,并采用长短期记忆(long short-term memory, LSTM)对行人进行轨迹预测; 文献[11]利用对比学习方法,提取行人动作信息,预测行人的多模态轨迹; 文献[12-13]虽然利用单目深度估计网络提取车辆的自我运动,但是没有考虑行人姿态对轨迹预测的影响.

1.3　多信息融合方法

将多种信息源的时间序列特征进行有效地融合是提升模型整体性能的关键,国内外学者针对于一维序列特征融合做了大量的相关研究. 文献[5]采用较为简单高效的拼接融合方法,将提取出的地图与行人的轨迹特征进行拼接,并采用门控循环单元(gated recurrent unit, GRU)进行特征解码; 文献[14]采用分级融合的方法,先将速度与轨迹序列特征进行拼接融合,融合后的特征与骨骼序列进行二次融合,采用注意力机制对最终的特征进行加权; 文献[15]提出一种多层次混合模型将文本、音频、视频的序列特征利用注意力机制进行融合; 文献[16]提出了一种基于transformer的融合方法,利用transformer使一个模态从另一个模态接收信息,实现多模态融合. 但是随着场景的变化,不同的信息在预测过程中对预测结果产生的影响也是动态变化的,上述的融合方法在信息融合前没有考虑这一问题,在信息评价的基础上进行多信息融合能更好地适应高度动态变化的交通场景.

2 基于多信息融合的轨迹预测模型

驾驶视角下行人轨迹预测的主要任务为:通过观测0.5 s(15帧)行人历史运动轨迹,从而预测未来最多1.5 s(45帧)的运动轨迹. 本文提出的MIFNet网络结构如图1所示,相比于单信息模型,将行人姿态和局部光流信息考虑在内,通过轨迹编码器、姿态编码器、光流编码器进行特征编码,并针对这3种信息进行信息评价,为多信息特征融合提供信息间的权重分配. 同时,利用跨信息融合注意力机制将3种特征进行融合. 通过条件变分自编码器(CVAE)学习融合特征X_t和真实轨迹Y_t的潜在联系,得到潜变量Z的概率分布,实现多模态轨迹预测. 轨迹预测网络包括目标解码器和双向轨迹解码器. 下文介绍每个部分的细节.

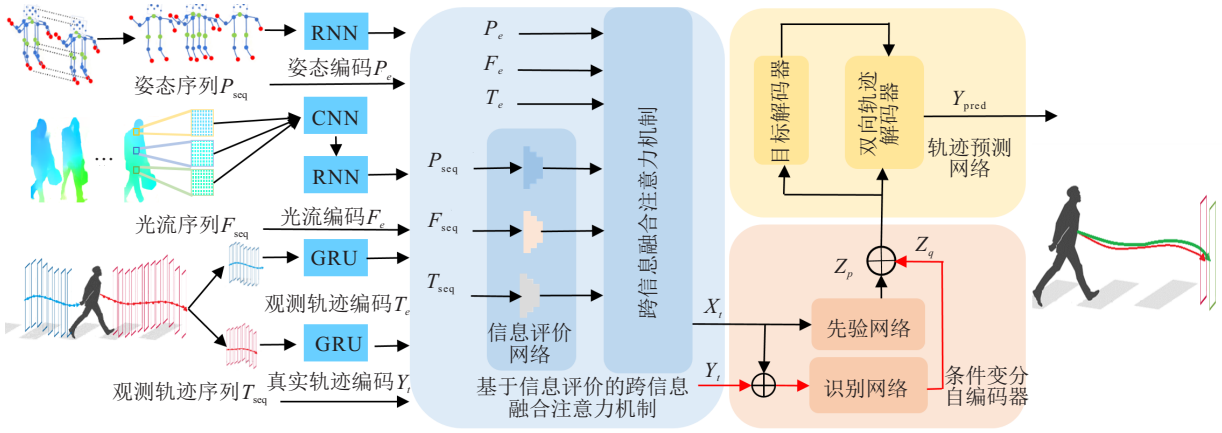


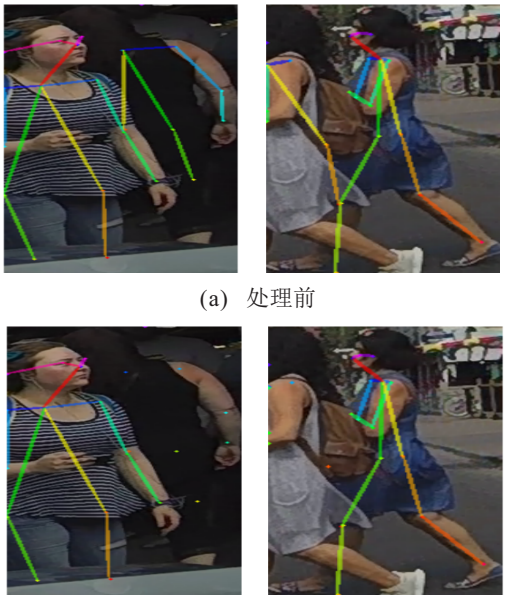
图 1 MIFNet网络结构

2.1 多信息编码网络

多信息编码网络包括姿态序列编码、光流序列编码、轨迹序列编码3部分,通过对姿态序列P_{seq}、光流序列F_{seq}和轨迹序列T_{seq}进行编码,分别得到编码后的特征P_e、F_e、T_e.

2.1.1 姿态编码器

对于行人i,假设他在某时刻的骨架向量为P_t^i,维度为18×2,则其观测的骨架信息可以表示为$P_{seq}^i = [P_{t-t_{obs+1}}^i, \dots, P_t^i]$,其中$t_{obs}$为观测时间. 本文采用2 D姿态估计网络OpenPose[17]提取行人骨骼关键点,OpenPose能同时对多人检测,但是在对行人轨迹进行估计时需要得到单人的姿态,当行人之间距离过近时,利用OpenPose很难提取到单个目标行人的骨架信息,会将周围行人的骨骼关键点一同输出,如图2(a)所示. 为了提取单个目标行人的骨架信息,过滤掉周围行人骨架信息的干扰,本文对OpenPose模型输出进一步处理,将检测到的行人骨骼关键点矩阵按照图片内行人个数划分为对应的单人骨骼关键点矩阵,分别计算单人关键点矩阵中检测到的关键点数目,数目最多的即为目标行人,处理结果如图2(b)所示.



(a) 处理前

(b) 处理后

图 2 OpenPose单人骨骼信息处理对比

由于行人骨骼关键点会因行人与行人、汽车、建筑物等遮挡造成残缺,从而导致姿态序列有较高的噪声,甚至会导致姿态序列失真难以表达行人意图. 针对这个问题,与文献[18]提出的骨骼关键点重建不同,本文提出一种基于关键帧的姿态序列重组方法,并采用时间序列建模能力较强的RNN网络对重组后的姿态序列编码,最终输出维度为1 × 32的姿态特征编码P_e. 观测的15帧时间较短(0.5 s),在此期间行人的姿态变化不大,步态动作较少,动作重复帧较多. 因此,保留序列的首帧和尾帧,在剩余帧中提取两帧作为关键帧,利用OpenPose估计的关键点置信度来评估某一帧的骨骼信息完整程度,取置信度和最高的两帧为关键帧. 将首尾帧和关键帧共4帧重组为新的姿

态序列. 采用重组的4帧能在包含姿态序列关键信息的同时过滤掉骨架信息遮挡、识别不全等问题的无效帧.

2.1.2　光流编码器

光流是指在连续的两帧图像中由于图像中的物体移动或者相机的移动导致的图像中目标像素的移动. 一方面,行人光流可以体现行人各部位更精确的运动信息,补充行人骨架信息评估行人姿态的不足,从而更好预测行人意图;另一方面,行人光流可以表征相机运动,从而表达车辆的自我运动信息. 对于行人i,假设他在某时刻t和上一时刻$t-1$的光流向量为F_t^i,则其观测的光流序列可以表示为$F_{\text{seq}}^i = [F_{t-t_{\text{obs}}}^i, \dots, F_t^i]$,其中$t_{\text{obs}}$为观测时间. 本文采用GMFlow[19]来估计行人的光流信息.

由于单帧的行人空间上光流变化范围不大,虽然包含了大量的光流数据但是包含信息较少,完整光流序列有较多无效信息,造成计算性能的浪费,同时也难以区分行人不同部位的有效特征. 此外,估计较远行人光流时同样会有遮挡问题,造成光流序列失真. 因此,本文设计一种基于CNN-RNN的局部光流特征提取网络,采用局部光流法,将行人分为头部、躯干、腿部$8\times8\times3$的光流区域;通过局部光流法提取行人各部位的中心区域光流,既能减少遮挡造成的影响,又能使行人头部、躯干、腿部运动区域光流对比更加明显,结合行人骨架信息更好地估计行人意图;对光流区域的光流信息进行极坐标映射和归一化处理得到光流序列F_{seq};利用卷积神经网络(convolutional neural network, CNN)分别提取观测序列中行人空间上的光流的特征,得到维度为14×24的头部、躯干、腿部光流特征,RNN网络擅长处理时序信息,利用RNN在时间维度提取卷积后的光流特征,CNN和RNN结合共同提取行人光流的时空特征,得到最终的局部光流编码F_e.

2.1.3　轨迹编码器

对于行人i,假设他在某时刻的观测位置向量为$T_t^i = (x_t^i, y_t^i, w_t^i, h_t^i)$. 其中:$(x_t^i, y_t^i)$表示行人$i$在$t$时刻的边界框中心点坐标,$(w_t^i, h_t^i)$表示为$t$时刻边界框的大小信息. 观测轨迹序列可以表示为$T_{\text{seq}}^i = [T_{t-t_{\text{obs}}+1}^i, \dots, T_t^i]$,预测轨迹为$\hat{Y}_{\text{pred}}^i = [Y_{t+1}^i, Y_{t+2}^i \cdots, Y_{t+\text{pred}}^i]$. 其中:$t_{\text{obs}}$为观测时间,$t_{\text{pred}}$为预测时间. 利用GRU对观测15帧的轨迹信息特征编码,得到观测轨迹编码特征T_e. 在训练过程中,为了训练CVAE,需要对真实轨迹Y_t利用GRU进行特征编码,得到真实轨迹编码特征Y_t.

2.2　基于信息评价的跨信息融合网络

在轨迹预测过程中,输入模型的3种信息源序列为观测轨迹信息、光流信息、姿态信息. 交通场景下,这3种信息源序列在不同的行人位置发挥作用的程度不同. 例如:当行人距离观测相机较远时,较难提取到完整的行人骨架信息,此时骨架信息表征的姿态信息会失真,从而影响轨迹预测效果;当行人距离观测相机较近时,轨迹坐标在图像中变化幅度较大,难以从观测轨迹序列预测稳定准确的未来运动,但此时骨架信息完整,动作清晰,通过骨架信息能较好反映行人的意图,从而更精准地预测未来轨迹. 对于光流信息,当行人较远时,可能受较近的车辆、行人等遮挡,从而不能精准获取行人的局部光流,导致光流信息失真.

针对以上问题,本文提出一种基于信息评价的跨信息融合网络,通过建立多源信息评价网络,对不同来源的信息序列进行数据质量评价,在同一时刻轨迹预测过程中对不同信息源的特征分配不同权重,使数据质量高的发挥更大的作用;同时,改进自注意力机制,引入残差连接,设计跨信息融合网络,结合评价权重的跨信息融合网络,能够更好地应对多行人、场景复杂的城市交通环境下的行人轨迹预测. 基于信息评价的跨信息融合注意力机制网络结构如图3所示,网络的输入为各个信息的序列信息和编码特征,输出为融合后的编码特征X_t.

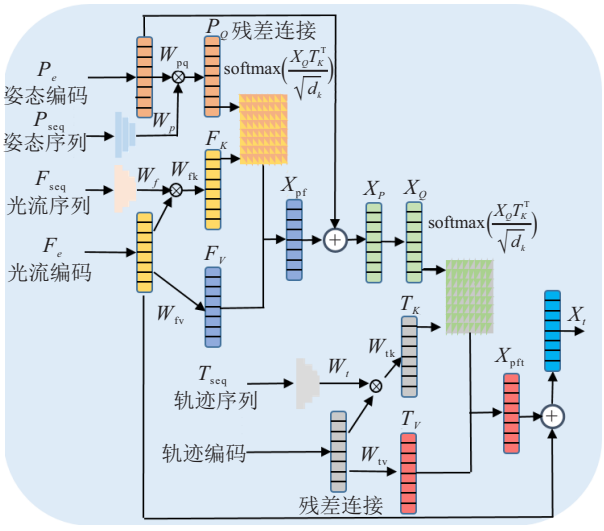


图 3　基于信息评价的跨信息融合注意力机制

利用信息评价网络得到各个序列信息的评价权重. 信息评价权重能够提高不同信息间更为重要的信息特征的敏感度,同时注意力机制能够聚焦单一信息特征内更为关键的特征,提升行人轨迹预测任务的效率和准确性. 信息评价的具体过程为:将获取的姿态序列P_{seq}、光流序列F_{seq}和轨迹序列T_{seq}分别输入

到前两层激活函数为ReLU、最后一层激活函数为Sigmoid的3层感知器中,得到对应的分配权重W_p、W_f、W_t,计算公式如下:

$$W_p = f_1(P_{\text{seq}}, W_1), \tag{1}$$

$$W_f = f_2(F_{\text{seq}}, W_2), \tag{2}$$

$$W_t = f_3(T_{\text{seq}}, W_3). \tag{3}$$

其中:W_p、W_f、W_t分别为姿态序列、光流序列、轨迹序列的数据评价权重,$f_1 \sim f_3$为感知器映射函数,P_{seq}、F_{seq}、T_{seq}分别为姿态、光流和轨迹序列,$W_1 \sim W_3$为感知机网络中总的可学习参数矩阵.

跨信息融合注意力网络引入ResNet网络[20]中残差的连接,在两两特征融合过程中加入残差信息,防止梯度消失,避免关键特征信息丢失. 注意力机制为利用transformer[21]中放缩的点积注意力改进得到的级联交叉注意机制. 在进行多信息融合时,首先对姿态特征和光流特征进行融合,姿态编码特征和姿态评价权重相乘,经权重为W_{pq}的线性变换得到新的特征P_Q. 光流编码有两个分支,其中一条结合评价权重W_f进行线性变换得到F_K,另一条通过可学习矩阵W_{fv}进行线性变换得到F_V,计算公式如下:

$$P_Q = P_e W_p W_{\text{pq}}^{\text{T}} + b_1, \tag{4}$$

$$F_K = F_e W_f W_{\text{fk}}^{\text{T}} + b_2, \tag{5}$$

$$F_V = F_e W_{\text{fv}}^{\text{T}} + b_3. \tag{6}$$

将P_Q和F_K点乘计算得到的注意力得分矩阵和F_V点乘,得到姿态和光流的融合特征X_{pf},X_{pf}和姿态编码特征进行残差连接,得到初步融合后的编码特征X_P. 计算公式如下:

$$X_P = \text{softmax}\left(\frac{P_Q F_K}{\sqrt{d_k}}\right) F_V + P_e. \tag{7}$$

轨迹编码特征的融合方法与姿态、光流融合方法类似,同样有两个分支,其中一条结合评价权重进行线性变换后得到T_K,初步融合后的特征X_P经过线性变换得到的X_Q,二者经过softmax计算注意力分数;另一条通过可学习矩阵W_{tv}线性变换后得到T_V,将T_V点乘注意力分数得到融合特征X_{pft}. X_{pft}和光流序列编码F_e进行残差连接得到最终的融合特征X_t. 计算过程如下:

$$T_K = T_e W_t W_{\text{tk}}^{\text{T}} + b_4, \tag{8}$$

$$X_Q = X_P W_{\text{tq}}^{\text{T}} + b_5, \tag{9}$$

$$T_V = T_e W_{\text{tv}}^{\text{T}} + b_6, \tag{10}$$

$$X_t = \text{softmax}\left(\frac{X_Q T_K}{\sqrt{d_k}}\right) T_V + F_e. \tag{11}$$

其中:P_e、F_e、T_e为编码器输出的姿态、光流和轨迹的编码特征,W_p、W_f、W_t为姿态、光流和轨迹序列的评价权重,W_{pq}、W_{fq}、W_{fv}、W_{tq}、W_{tk}、W_{tv}为线性变换的参数矩阵,$b_1 \sim b_6$为线性变化的偏差,P_Q、F_K、F_V、T_K、T_V、X_{pf}、X_P、X_Q、X_{pft}为线性变换得到的中间特征向量,$\sqrt{d_k}$为一个用来放缩的比例因子.

2.3 条件变分自编码器

本文借鉴CAVE模型[22]和BitaP模型[7]实现多模态目标和轨迹预测,网络结构如图4所示. 其中:虚线路径表示训练过程,实线路径表示推理过程. CVAE包含两个子模块:条件先验网络$p_\theta(Z|T_{\text{seq}})$,用于根据观测值对潜在变量$Z$进行建模;识别网络$q_\phi(Z|T_{\text{seq}}, Y_{\text{seq}})$,用来捕获$Z$与$Y_{\text{seq}}$之间的依赖关系. 采用非参数模型,不假设目标$Y_{\text{seq}}$的分布,而是通过学习$Z$的分布隐式地学习它. CVAE的目标是最大化其变分下界,如下所示:

$$\max_{\theta,\phi,\varphi} E_{q_\phi(Z|T_{\text{seq}},Y_{\text{seq}})}(\log_{p_\phi}(Y_{\text{seq}}|T_{\text{seq}}, Z)) - \text{KL}(q_\phi(Z|T_{\text{seq}}, Y_{\text{seq}})||p_\theta(Z|T_{\text{seq}})). \tag{12}$$

其中:第1项使预测分布中目标的对数似然期望最大化,第2项KL散度使先验网络和识别网络的分布差异最小化,T_{seq}、Y_{seq}、Z分别为观测轨迹序列、真实未来轨迹序列和潜变量.

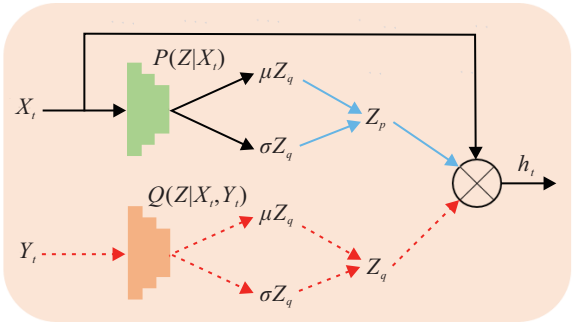


图 4 条件变分自编码器网络结构

2.4 轨迹预测网络

轨迹预测网络结构如图5所示,可以分为两部分:目标解码网络$p_\omega(G_t|X_t, Z)$、轨迹解码网络$p_\psi(Y_t|X_t, G_t, Z)$. 利用条件变分自编码器输出的h_t,目标解码网络$p_\omega(G_t|X_t, Z)$可以预测多模态目标G_t. 目标解码网络采用层数为3的感知机模型,预测目标G_t用作双向轨迹生成网络的输入,双向轨迹解码网络包含前向GRV和反向GRU. 前向GRU类似于常规GRU解码器,反向GRU从编码器隐藏状态h_t初始化,它将估计目标作为初始输入,并从时间$t+\delta$传播到$t+1$,因此向后的隐藏状态从目标更新到当前位置. 将同一时间步的前向和反向隐藏状态串联起来,以预测最终的行人运动轨迹[7,22]. 这些步骤可以表述为

$$h_{t+1}^{g}=\mathrm{GRU}_{g}(h_{t}^{g},W_{g}^{i}h_{t}^{g}+b_{g}^{i}),\tag{13}$$

$$h_{t+\delta-1}^{b}=\mathrm{GRU}_{f}(h_{t+\delta}^{b},W_{b}^{i}\hat{Y}_{t+\delta}^{g}+b_{b}^{i}),\tag{14}$$

$$\hat{Y}_{t+\delta-1}^{b}=W_{g}^{o}h_{t+\delta-1}^{f}+W_{b}^{o}h_{t+\delta-1}^{b}+b^{o}.\tag{15}$$

其中: g、b、i、o分别代表前向传播、反向传播、输入和输出, h_{t+1}^{g}和$h_{t+\delta}^{b}$表示通过两个不同的网络传递h_t.

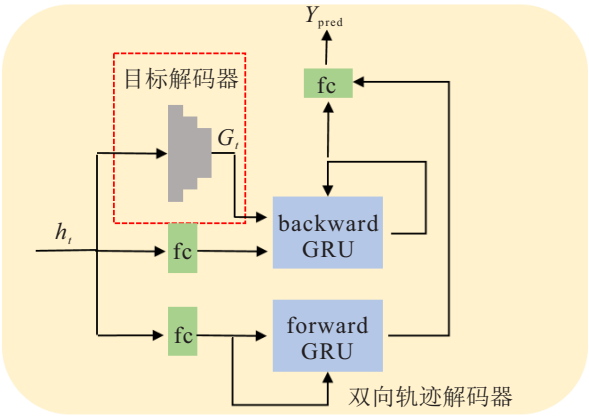


图 5　轨迹预测网络

2.5　损失函数

本文引入目标预测, 与Bitap[7]模型类似, MIFNet的损失函数包括目标损失、轨迹损失和先验网络分布与识别网络分布之间的损失, 如下所示:

$$L=\min_{i\in N}\|G_t-T_{\mathrm{loc}}-\hat{G}_t^i\|+\min_{\tau=t+1}^{t+\delta}\|Y_\tau-T_{\mathrm{loc}}-\hat{Y}_\tau^i\|+\mathrm{KLD}.\tag{16}$$

其中: G_t和Y_τ分别为行人真实目标和真实轨迹, $\hat{G}_t$和$\hat{Y}_\tau$分别为依据观测轨迹信息预测出来的远期目标和未来轨迹, T_{loc}为行人当前位置, KLD为先验分布和判别分布之间的KL散度损失.

3　实验结果与分析

3.1　实验数据集和评价指标

本文采用JAAD和PIE数据集来评估MIFNet预测模型. JAAD和PIE包括车载相机录制的交通场景下行人运动视频, 其中JAAD包含2 800条轨迹, PIE包含1 835条轨迹. 所有视频均以每秒30帧的速度捕获, 1 920 × 1 080像素. 本文遵循前人工作[7,9-11], 丢弃长度小于2 s的轨迹, 并使用0.5 s的观测数据作为输入, 以预测长度为0.5 s、1.0 s和1.5 s的未来轨迹.

MIFNet的主要评估指标包括平均位移误差(average displacement error, ADE)和最终位移误差(final displacement error, FDE), 分别用来测量整个轨迹的精度和轨迹终点的精度. 计算过程使用均方误差(mean square error, MSE)[7,9-11]来评估本文模型在JAAD和PIE上的表现, 根据边界框的左上和右下坐标计算ADE和FDE. 采用中心均方误差(central mean square error, CMSE)和中心最终均方误差(central final mean square error, CFMSE)来计算中心点的ADE和FDE. 这两个误差度量与MSE类似, 不过它们是基于边界框中心计算的.

3.2　实验设置

MIFNet中的姿态和光流编码器的隐藏层维度为32, 轨迹编码和解码器的隐含层维度为256. 所有模型均在Ubuntu 20.04系统、NVIDIA 3080显卡上进行训练和测试, 训练批次大小为128, 学习率为0.001.

本文对比的基线模型包括: 早期的轨迹预测模型, 如线性卡尔曼滤波器、长短期记忆模型(LSTM)、贝叶斯LSTM模型(B-LSTM)、FOL-X; 近期的预测模型, 如$\mathrm{PIE}_{\mathrm{traj}}$、BtriP模型以及当前state-of-the-arts (SOTA)模型SGNet和ABC+[7,9-11,23].

3.3　实验结果与分析

3.3.1　MIFNet与基线模型结果比较

MIFNet通过观察行人在视频中的历史运动轨迹, 预测出行人未来的多条运动轨迹, 并计算最佳轨迹和真实轨迹像素级别平均位移和最终位移的均方误差. 同时, 为了与以往单条轨迹预测模型进行对比, 将CVAE模块去掉, 得到模型MIFNet-D. 为了后续方便消融实验, 将结合姿态、光流和轨迹信息的模型命名为MIFNet-FP. 表1显示了MIFNet模型和基线模型的对比结果, 从结果上看: MIFNet-FP在PIE数据集上0.5 s的ADE和1.5 s的CADE均优于已有的预测模型, 且其他指标与最优结果相近; 在JAAD数据集上1.5 s长时轨迹预测的任务中取得了最佳预测结果, 优于已有的SOAT模型SGNet和ABC+, 在0.5 s和1.0 s的预测任务中与SOAT模型预测误差比较接近.

表2中比较了主流模型的参数量和在整个测试集的推理时间. 所有模型均在相同的硬件环境下进行测试, 推理时间为采用相同的批次大小(1 024)推理整个测试集行人轨迹所用时长, 并多次测量取平均值. 本文模型MIFNet-FP在参数量上比BiTraP-NP略有增加, 但在JAAD和PIE测试集的推理时间增加不大, 分别增加了1.82 s和3.04 s. MIFNet相比于当前SOAT模型SGNet-ED减少了52.49 %的参数量, 并且推理速度大幅度增加: 在JAAD测试集上推理时间减少了84.44 %, 在PIE测试集上推理时间减少了85.79 %.

表 1 MIFNet与基线模型结果比较

方法	JAAD数据集					PIE数据集				
	ADE (0.5 s)	ADE (1.0 s)	ADE (1.5 s)	CADE (1.5 s)	CFDE (1.5 s)	ADE (0.5 s)	ADE (1.0 s)	ADE (1.5 s)	CADE (1.5 s)	CFDE (1.5 s)
Linear	233	857	2 303	1 565	6 111	123	477	1 365	950	3 983
LSTM	289	569	1 558	1 573	5 766	172	330	911	837	3 352
B-LSTM	159	539	1 535	1 447	5 615	101	296	855	811	3 259
FOL-X	147	484	1 374	1 290	4 925	47	183	584	546	2 303
PIE_{traj}	110	399	1 280	1 183	4 780	58	200	636	596	2 477
PIE_{full}	–	–	–	–	–	–	–	556	520	2 162
BiTraP-D	93	378	1 206	1 105	4 565	41	161	511	481	1 949
MIFNet-D	84	348	1 110	1 062	4 294	36	141	459	431	1 843
BiTraP-NP (20)	38	94	222	177	565	23	48	102	81	261
SGNet-ED (20)	**37**	**86**	197	146	443	16	39	88	66	206
ABC+ (20)	40	89	189	145	409	16	**37**	**87**	65	**191**
MIFNet-FP (20)	41	89	**187**	**135**	**382**	**16**	38	88	**63**	192

表 2 不同模型的参数量和推理时间

方法	参数量 / 个	JAAD测试集推理速度 / s	PIE测试集推理速度 / s
PIE_{full}	1 240 372	–	2.99
BiTraP-NP (20)	1 538 024	5.48	13.63
MIFNet-FP (20)	3 618 101	7.30	16.67
SGNet-ED (20)	7 622 406	45.47	117.32

虽然JAAD数据集上MIFNet-FP在0.5 s和1.0 s的轨迹预测任务中位移误差较SGNet-ED略大,但是MIFNet-FP的优势在于,1.5 s长时轨迹预测任务中位移误差最小并具有更小的参数量和更快的推理速度. 一般而言,长时轨迹在自动驾驶领域更有意义,能使自动驾驶汽车进行路径规划时更早地避让行人.

3.3.2 消融实验

为了验证基于置信度信息的姿态关键帧提取的合理性,以及关键帧选取的最优帧数,本文以完整骨架序列模型为基准,对比关键帧为1、2、4、6、8、13帧时的模型预测结果,如表3所示. 其中关键帧13帧为基线模型,其为完整的骨架序列去掉首尾帧后的帧数. 从表3可以看出关键帧数目不同会影响模性预测性能. 当关键帧数目为1帧时,其预测效果与全序列相比没有明显优势,取1帧为关键帧的重组序列帧数较少,不能较完整地包含行人运动过程中的关键姿态信息. 当关键帧数目较多时(如8帧),其预测效果同样没有明显优于关键帧为13的全序列预测结果,原因可能为没有较好地过滤掉出现行人遮挡的无效帧. 当关键帧数为2时,重组的骨架序列帧在包含行人姿态关键信息的同时,过滤掉了遮挡的无效帧,其在JAAD和PIE数据集的整体预测结果明显优于其他帧数: JAAD数据集上1.5 s的ADE、CADE和CFDE分别较全序列减少5.31 %、5.77 %和10.82 %, PIE数据集上1.5 s的ADE、CADE和CFDE分别较全序列减少3.16 %、5.63 %和6.04 %.

表 3 不同关键帧数目消融实验结果

关键帧数目	JAAD数据集					PIE数据集				
	ADE (0.5 s)	ADE (1.0 s)	ADE (1.5 s)	CADE (1.5 s)	CFDE (1.5 s)	ADE (0.5 s)	ADE (1.0 s)	ADE (1.5 s)	CADE (1.5 s)	CFDE (1.5 s)
1	44	97	205	154	440	18	43	102	78	253
2	**42**	**93**	**196**	**147**	415	18	**41**	**92**	**67**	**215**
4	44	97	204	149	**404**	18	42	96	73	230
6	**42**	95	207	154	453	18	42	100	76	245
8	43	95	206	157	449	18	42	98	72	247
13	43	97	207	156	453	18	42	95	71	228

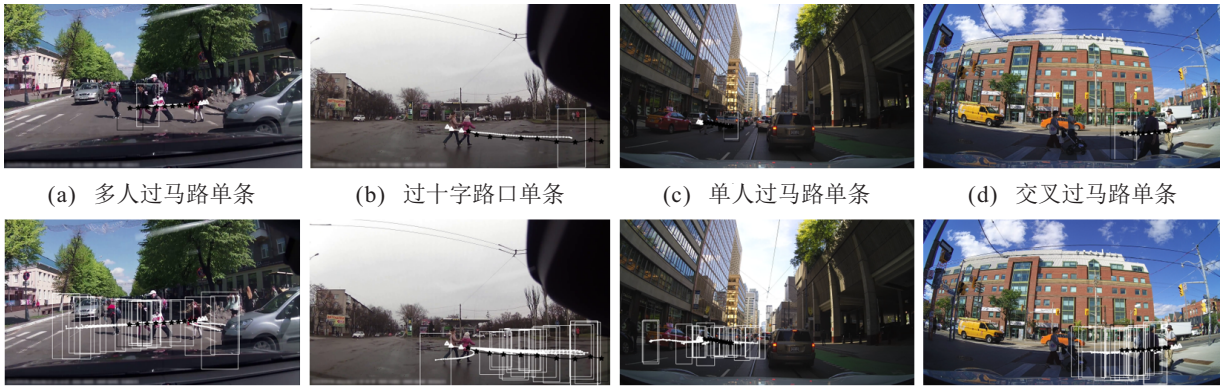
表4显示了不同模块的消融实验,其中BiTraP-D、MIFNet-D为去掉生成概率模型网络CAVE后的模型,此时模型变为单条轨迹输出. 从表4可以看出,单条轨迹预测模型在JAAD数据集和PIE数据集上预测效果较BiTraP-D有明显提升. MIFNet-P (20)为结合轨迹信息和姿态信息并生成20条预测轨迹的网络模型, MIFNet-F (20)为结合轨迹信息和光流信息的网络模型,可以看出,在轨迹信息的基础上引入姿态信息或光流信息均能减少模型预测位移误差. MIFNet-PJ (20)为将姿态、光流和轨迹特征进行直接拼接的网络模型, MIFNet-FP (20)为采用信息评价的跨信息融合注意力机制融合多信息特征的网络模型,可以看出,本文提出的基于信息评价的多信息融合注意力机制的融合效果显著好于多信息特征直接拼接的效果,并且融合3个信息特征的模型明显好于只融合两个信息特征的模型.

表 4　MIFNet不同模块的消融实验

方法	JAAD数据集					PIE数据集				
	ADE (0.5 s)	ADE (1.0 s)	ADE (1.5 s)	CADE (1.5 s)	CFDE (1.5 s)	ADE (0.5 s)	ADE (1.0 s)	ADE (1.5 s)	CADE (1.5 s)	CFDE (1.5 s)
MIFNet-D	84	348	1 110	1 062	4 294	36	141	459	431	1 843
BiTraP-NP (20)	38	94	222	177	565	23	48	102	81	261
MIFNet-P (20)	42	93	196	147	415	18	41	92	67	215
MIFNet-F (20)	41	91	194	146	408	18	43	100	76	256
MIFNet-PJ (20)	42	95	203	153	432	21	46	100	73	225
MIFNet-FP (20)	41	89	187	135	382	16	38	88	63	192

3.3.3　模型预测轨迹结果定性分析

图6显示了MIFNet模型在测试集上的预测结果, 图6(a)和6(b)为JAAD数据集上单条轨迹预测结果, 图6(c)和6(d)为PIE数据集上单条轨迹预测结果, 图6(e)～6(h)为对应的多条轨迹预测结果. 预测结果在观测最后一帧上绘制, 行人未来轨迹中, 稀疏黑色五角星串联的为行人真实轨迹, 稠密白色圆点串联的为模型预测的行人轨迹. 在单条轨迹的预测中, MIFNet模型在预测终点位置和行人框大小时, 相当于真实值偏差较大, 但整体轨迹点运动趋势和运动曲线预测相对较准. 由于使用了生成概率预测模型, MIFNet能预测未来行人运动的20条运动轨迹, 预测轨迹的密集程度反映行人的运动趋势, 如图6(f)中19条预测轨迹显示行人将会朝前过马路, 仅有一条预测结果为行人突然掉头转向, 整体趋势为朝前过马路, 其与真实行人运动趋势一致. 此外, 在多条运动轨迹预测中, 其预测的轨迹和行人大小基本覆盖了真实轨迹行人大小, 综合考虑了未来行人运动轨迹的所有可能, 使自动驾驶汽车决策时能更全面地进行路径规划, 提升自动驾驶汽车行驶的安全性.



(a) 多人过马路单条　(b) 过十字路口单条　(c) 单人过马路单条　(d) 交叉过马路单条

(e) 多人过马路多条　(f) 过十字路口多条　(g) 单人过马路多条　(h) 交叉过马路多条

图 6　JAAD和PIE数据集下的轨迹预测结果

4　结　论

多信息融合的行人轨迹预测, 相比于单信息的轨迹预测, 能更精准地预测行人的远期目标和意图, 从而降低模型预测过程中的预测位移误差. 本文针对行人遮挡造成骨骼关键点识别残缺问题, 以单帧置信度和的大小为依据提取骨架序列关键帧, 进而重组骨架序列. 通过引入行人光流变化信息表征相机的相对运动, 在光流特征编码时重点考虑局部光流信息. 在多信息特征融合过程中, 依据信息评价权重, 在跨信息融合注意力机制的作用下融合多信息特征. 本文提出的预测模型MIFNet在JAAD和PIE数据集上相较于其他预测方法在位移误差方面有一定程度的提升, 在1.5 s的长时轨迹预测任务中以较少的参数量、较快的推理速度取得了最佳的行人轨迹预测效果. 在JAAD数据集上0.5 s和1.0 s的行人轨迹预测精度上还有提升的空间, 后续将针对这个薄弱环节进一步研究.

参考文献(References)

[1] 孔玮, 刘云, 李辉, 等. 基于深度学习的行人轨迹预测方法综述[J]. 控制与决策, 2021, 36(12): 2841-2850. (Kong W, Liu Y, Li H, et al. Survey of pedestrian trajectory prediction methods based on deep learning[J]. Control and Decision, 2021, 36(12): 2841-2850.)

[2] Yagi T, Mangalam K, Yonetani R, et al. Future person localization in first-person videos[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7593-7602.

[3] Gupta A, Johnson J, Li F F, et al. Social GAN: Socially

acceptable trajectories with generative adversarial networks[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 2255-2264.

[4] Gu T P, Chen G Y, Li J L, et al. Stochastic trajectory prediction via motion indeterminacy diffusion[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 17092-17101.

[5] Salzmann T, Ivanovic B, Chakravarty P, et al. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data[C]. European Conference on Computer Vision. Cham: Springer, 2020: 683-700.

[6] Mangalam K, An Y, Girase H, et al. From goals, waypoints & paths to long term human trajectory forecasting[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 15213-15222.

[7] Yao Y, Atkins E, Johnson-Roberson M, et al. BiTraP: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1463-1470.

[8] 郭戈, 胡峻豪. 区别多种出行方式的城市活动轨迹预测[J]. 控制与决策, 2023, 38(4): 1022-1030.
(Guo G, Hu J H. Urban activity trajectory prediction with different travel modes[J]. Control and Decision, 2023, 38(4): 1022-1030.)

[9] Wang C H, Wang Y C, Xu M Z, et al. Stepwise goal-driven networks for trajectory prediction[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2716-2723.

[10] Rasouli A, Kotseruba I, Kunic T, et al. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction[C]. 2019 IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 6261-6270.

[11] Halawa M, Hellwich O, Bideau P. Action-based contrastive learning for trajectory prediction[C]. European Conference on Computer Vision. Cham: Springer, 2022: 143-159.

[12] 汪梓豪, 蔡英凤, 王海, 等. 基于单目视觉运动估计的周边多目标轨迹预测方法[J]. 汽车工程, 2022, 44(9): 1318-1326.
(Wang Z H, Cai Y F, Wang H, et al. Surrounding multi-target trajectory prediction method based on monocular visual motion estimation[J]. Automotive Engineering, 2022, 44(9): 1318-1326.)

[13] Neumann L, Vedaldi A. Pedestrian and ego-vehicle trajectory prediction from monocular camera[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 10199-10207.

[14] Yang D F, Zhang H L, Yurtsever E, et al. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention[J]. IEEE Transactions on Intelligent Vehicles, 2022, 7(2): 221-230.

[15] 宋云峰, 任鸽, 杨勇, 等. 基于注意力的多层次混合融合的多任务多模态情感分析[J]. 计算机应用研究, 2022, 39(3): 716-720.
(Song Y F, Ren G, Yang Y, et al. Multimodal sentiment analysis based on hybrid feature fusion of multi-level attention mechanism and multi-task learning[J]. Application Research of Computers, 2022, 39(3): 716-720.)

[16] Tsai Y H H, Bai S J, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[J]. Proceedings of the Conference Association for Computational Linguistics Meeting, 2019, 2019: 6558-6569.

[17] Cao Z, Simon T, Wei S H, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1302-1310.

[18] Heidari N, Iosifidis A. Progressive spatio-temporal graph convolutional network for skeleton-based human action recognition[C]. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, 2021: 3220-3224.

[19] Xu H F, Zhang J, Cai J F, et al. GMFlow: Learning optical flow via global matching[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 8111-8120.

[20] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Conference and Workshop on Neural Information Processing Systems. California: MIT Press, 2017: 5998-6008.

[22] Sohn K, Yan X C, Lee H. Learning structured output representation using deep conditional generative models[C]. Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015: 3483-3491.

[23] Du X X, Vasudevan R, Johnson-Roberson M. Bio-LSTM: A biomechanically inspired recurrent neural network for 3-D pedestrian pose and gait prediction[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 1501-1508.

作者简介

桑海峰(1978−), 男, 教授, 博士生导师, 从事机器视觉检测和智能视频分析等研究, E-mail: sanghaif@163.com;

刘泉恺(1998−), 男, 硕士生, 从事行人轨迹预测的研究, E-mail: liuqk_sut@163.com;

王金玉(1996−), 女, 博士生, 从事行人轨迹预测的研究, E-mail: 1911131982@qq.com;

陈旺兴(1998−), 男, 博士生, 从事行人轨迹预测的研究, E-mail: 1909703861@qq.com.