

基于图卷积的3D骨架数据的双人交互行为识别

张静亭¹, 曹江涛¹, 姬晓飞²

(1. 辽宁石油化工大学 信息与控制工程学院, 辽宁 抚顺 113001; 2. 沈阳航空航天大学 自动化学院, 辽宁 沈阳 110136)

摘 要: 针对图卷积神经网络的双人交互行为识别方法存在交互语义信息表达不充分的问题, 提出了一种新的双人交互时空图卷积神经网络(DHI-STGCN)用于行为识别的方法。该网络包含空间子网络模块和时间子网络模块。将基于交互动作视频获取的3D骨架数据生成一种双人交互动作的空间动作图用于空间信息的表示, 图中根据关节点位置信息对双人之间的连接边赋予不同的权重。时间信息处理中, 在构造的邻接矩阵中增加了上下文时间信息的联系, 图中关节点与其一定时间范围内的节点增加连接。将生成的时空图数据送入空间图卷积网络模块, 结合时间图卷积网络模块增强帧间运动特征连续性进行时序建模。该模型充分考虑了双人交互动作的紧密关系, 具有较强的鲁棒性, 获得了比现有模型更好的交互动作识别效果。

关键词: 时空图卷积; 骨架数据; 双人交互; 行为识别

中图分类号: TP391.1

文献标志码: A

doi:10.12422/j.issn.1672-6952.2023.03.014

3D Skeleton Data Double Human Interaction Recognition Based on Graph Convolution Network

Zhang Jingting¹, Cao Jiangtao¹, Ji Xiaofei²

(1. School of Information and Control Engineering, Liaoning Petrochemical University, Fushun Liaoning 113001, China;

2. School of Automation, Shenyang Aerospace University, Shenyang Liaoning 110136, China)

Abstract: Aiming at the problem of insufficient representation of interactive semantic information in the double human interaction behavior recognition method based on graph convolutional neural networks, a new double human interactive spatial-temporal graph convolution network (DHI-STGCN) was proposed for behavior recognition. The network contains spatial sub-network modules and temporal sub-network modules. Based on the 3D skeleton data obtained from the interactive action video, a spatial action graph of double human interactive action was generated for the representation of spatial information. In the graph, the connecting edges between double human were given different weights according to the joint point position information. The connection of context time information was added in the constructed adjacency matrix, and the joint points in the graph were connected with their nodes within a certain time range in time information processing. The generated spatial-temporal graph data was sent to the spatial graph convolution network module, and the temporal graph convolution network module was combined to enhance the continuity of inter frame motion features for modeling in time. The model fully considers the close relationship of double human interaction. The comparative experimental results on NTU-RGB+D dataset show that the algorithm has strong robustness and obtains better interaction recognition effect than the existing models.

Keywords: Spatial-temporal graph convolution; Skeleton data; Double human interaction; Behavior recognition

基于视频的人体行为识别技术在智能安防^[1]、运动分析^[2]、手势识别^[3]、交互动作识别^[4]等领域具有非常广阔的应用前景。随着微软Kinect相机的迅速普及, 很多大型基于视频的关节点数据集的引入为研究者们进行人类行为识别的研究提供了优越条

件。针对人体骨架数据的行为识别的主要任务是动作分类和识别, 基于骨架的动作识别由于其对动态的环境和复杂的背景有很强的适应性被广泛地关注和研究。

基于卷积神经网络(Convolutional Neural

收稿日期: 2021-12-16 修回日期: 2022-01-21

基金项目: 国家自然科学基金项目(61673199)。

作者简介: 张静亭(1995-), 女, 硕士研究生, 从事基于深度学习的行为识别方面研究; E-mail: 18946104650@163.com。

通信联系人: 曹江涛(1978-), 男, 博士, 教授, 博士生导师, 从事智能方法及其应用、视频分析与处理方面研究; E-mail: cigroup@126.com。

Network, CNN)的方法通常手动设计转换规则。Q. Ke等^[5]利用人体关节序列,将每个关节生成一个向量,通过空间结构转换为二维的灰度图像,再送入CNN中进行动作。这种CNN方法并不能很好地对双人交互行为数据进行时序建模来表征双人交互的关系。

循环神经网络(Recurrent Neural Networks, RNN)可以对双人骨架数据的时间信息进行建模,然而RNN只能表示相邻较近节点的依赖关系。因此,J.Liu等^[6]提出了一种将长短时记忆网络(Long Short Term Memory, LSTM)扩展到时空域的时空长短时记忆网络(Spatial-Temporal Long Short-Term Memory, ST-LSTM)。I.Lee等^[7]提出了一种包含短期、中期和长期集成的长短时记忆网络(Temporal Sliding LSTM, TS-LSTM),先将骨架数据缩放、旋转、平移变换到另一个坐标系,再从变换的数据中提取运动特征。S.Li等^[8]提出了一种新的可独立循环的神经网络(Independent Recurrent Neural Network, IndRNN),并用于双人交互行为识别。该网络中,同一网络层的神经元组织相互独立,在不同网络层上相互关联,有效地减弱了梯度爆炸和梯度消失。P.F.Zhang等^[9]在RNN结构中对相对重要的因素加入了注意力机制(Element-Wise Attention Gate, EleAttG),增加了网络的输入对输出的贡献。但是,RNN往往侧重于对时间信息的表示,而忽略人体运动中的空间结构信息^[10]。

基于图卷积网络(Graph Convolutional Network, GCN)的方法能够表示点和连线的关系,在基于人体3D骨架数据的动作识别研究中取得了诸多成果。S.Yan等^[11]构建时空图卷积网络(Spatial-Temporal Graph Convolutional Network, ST-GCN),构造了人体关节的自然连接空间图。Y.H.Wen等^[12]提出了一种基于主题编码分层空间结构的图卷积网络,利用人体的骨架序列使用可变的时间密集块来获取不同范围的局部时间信息。Y.Tang等^[13]提出了一种深度渐进式强化学习(Deep Progressive Reinforcement Learning, DPRL)的帧选择方法,对最具信息量的帧进行选择,忽略信息不明确的帧。Y.H.Wen等^[12]在时序信息上提取了长范围和短范围的时间依赖信息,加入时域注意力机制,有效地融合了来自骨架连接和断开关节的不同语义信息。尽管基于GCN的方法在人体3D骨架数据上提供了很多行为识别方案,GCN使用的骨架空间图是仅依赖于人身体物理结构的预定义图,对

细微的交互动作识别不够充分,不能很好地对双人交互关系进行整体建模。因此,本文提出一种双人交互时空图卷积神经网络(Double Human Interaction Spatio-Temporal Graph Convolution Network, DHI-STGCN)的行为识别模型,在NTU-RGB+D大型数据集上进行了实验。结果表明,此模型进一步提升了双人交互行为识别的准确性。

1 算法框架

DHI-STGCN包含空间图卷积模块和时序图卷积模块。骨架数据中关节点坐标对应图的节点,而人身体结构的连通性及时间上的连通性对应图的两类边。将双人交互动作的骨架数据为模型的输入,先对每帧骨架数据生成的空间动作图进行空域卷积,将得到的结果作为特征值,再进行时域上的卷积,发掘更全面的双人交互动作时空信息,不再局限于单人的局部关节范围。DHI-STGCN算法框架如图1所示。

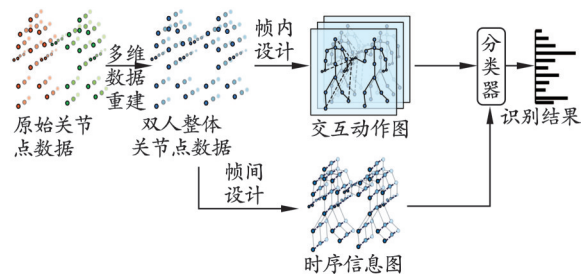


图1 DHI-STGCN 算法框架

2 时空建模

2.1 空间信息构图

图是一种通过顶点和连线描述的数据类型,空间动作图如图2所示。图中,实线表示构造的空间动作图包含骨架结构的连线;虚线表示没有结构关系的两个人的关节点间的联系。例如,握手的动作,一个人的手部关节会和另一人的手部关节发生近距离接触,但是他们并没有在骨骼结构中相连,手部节点距离很远,需要经过多层图卷积才能把信息传给对方,通过手动增加连接,建立交互关系。

一帧完整的双人交互动作空间图所有关节点的连接关系通过一个对称的加权矩阵 W 表示,其中包含的权重 w 通过关节点进行确定,其权值定义如下:

$$w = \begin{cases} w_{ij} = \begin{cases} 0, & i=j \\ 1, & i \neq j \end{cases} \\ w_{a_i, a_j} = \Omega, & \Omega \in (0, 1) \end{cases} \quad (1)$$

式中, w_{ij} 为第 i 个关节点与第 j 个关节点的连接权

值,当不存在连接关系时,对应的权值为0,当存在自然结构的连接关系时,对应的权值为1; w_{a_i, a_j} 为两

人之间具有交互关系的关节点对应的权值,表达交互行为的空间依赖关系。

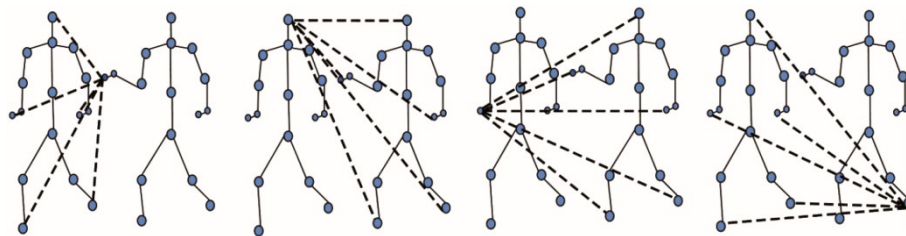


图2 空间动作图

这种权重的分配策略可以差异化地学习不同关节点的特征信息。利用双人骨架的空间结构表示交互行为的每帧动作信息,对交互动作的背景变化及场景的变化具有较好的适应性。

2.2 时序信息构图

在时间维度连接双人交互的时序信息图中,关节点与其周围存在具有时间依赖关系的关节点,提出了基于 n 阶时间结构的跨时间段邻接矩阵 A 来增强帧级运动特征连续性。 n 阶时间结构的邻接矩阵的关节点之间的连接为顺序连接,前节点与后 m 个邻居节点存在连接,时序信息如图3所示。为提取双人空间动作图的全局上下文时间信息,时域图卷积基于时序信息图与其邻接矩阵 A 实现局部上下文信息的时序建模,生成邻接矩阵表示上下文时间的关联信息,从而实现全局信息的时序信息建模。

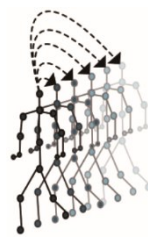


图3 时序信息

3 实验结果与分析

3.1 实验数据

为评估模型的识别准确性和泛化能力,实验使用大规模动作识别数据集NTU-RGB+D^[14],开展骨架数据的双人交互行为识别的研究。对双人NTU-RGB+D的骨架数据进行了三维可视化,结果如图4所示。其中,三个坐标轴是将双人骨架数据中的三维坐标值归一化后的相对坐标。

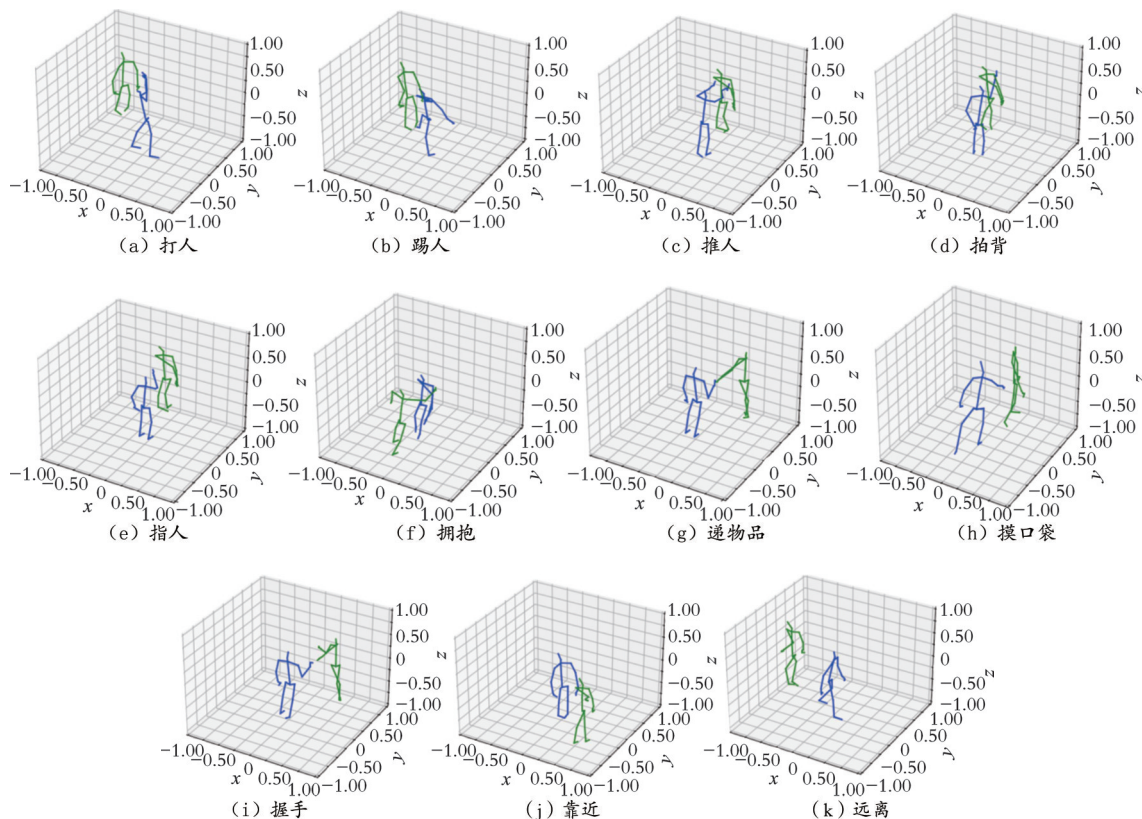


图4 双人NTU-RGB+D的骨架数据

NTU-RGB+D数据集包含10 428个双人交互动作视频,有11类交互动作。参与实验的人在室内使用三台不同位置的Kinect摄影机进行同时拍摄,通过摄影机内的深度传感器对每个人运动过程中的3D骨架数据位置进行记录。数据常用的实验划分方式有交叉受试者划分(Cross-Subject,CS)方式和交叉视角划分(Cross-View,CV)方式两种。此次对双人交互动作的骨架数据进行实验,采用交叉受试者划分的方法,进行了对比实验,由此检测双人交互时空图卷积神经网络的识别准确率。图卷积神经网络结构说明见表1。

表1 图卷积网络结构说明

层数	网络层	输入输出通道
1	SGCN-TGCN-0	(3,64)
2	SGCN-TGCN-1	(64,64)
3	SGCN-TGCN-2	(64,64)
4	SGCN-TGCN-3	(64,64)
5	SGCN-TGCN-4	(64,128)
6	SGCN-TGCN-5	(128,128)
7	SGCN-TGCN-6	(128,128)
8	SGCN-TGCN-7	(128,256)
9	SGCN-TGCN-8	(256,256)
10	SGCN-TGCN-9	(256,256)

表中,SGCN为空间图卷积模块,TGCN为时间图卷积模块。网络由10层时间图卷积层和10层空间图卷积层组成,前4个时空图卷积层具有64个输出通道,中间3个时空图卷积层有128个输出通道,最后3个时空图卷积层有256个输出通道。网络随机丢弃数据的概率设为0.5,以避免过拟合。其中,第5和第8图卷积层的步长为2,并将其作为采样层,通过全局池化的方式,得到256维的特征向量,最后经过SoftMax分类器进行分类识别。

NTU-RGB+D数据集的双人骨架数据样本统一帧数为300。本实验在Ubuntu16.04的操作系统下进行,基于python3.7软件的深度学习框架PyTorch开展,优化策略采用随机梯度下降法,设置初始的学习率衰减为0.1,共训练50轮,每批次训练的样本数为16个,利用交叉熵作梯度损失函数进行反向传播,权重的衰减系数为0.000 1,其他参数保持默认值。

3.2 测试结果与分析

在NTU-RGB+D数据集上进行了骨架数据的交互行为识别的实验。本模型的实验过程采用交叉视角CV划分方式,模型测试准确率如图5所示。

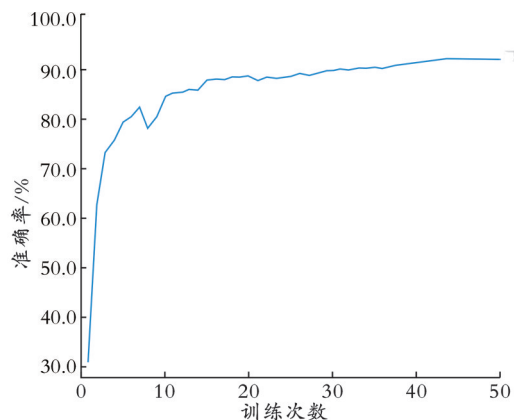


图5 模型测试准确率

从图5可以看出,模型的测试准确率曲线最终收敛,趋于平稳后得到的准确率为92.4%。为了进一步分析此模型的性能,生成如图6所示的混淆矩阵。图中,颜色从浅到深表示准确率从小到大,横坐标预测标签表示测试数据的动作类型,纵坐标真实标签表示识别的动作结果,数字表示识别的准确率。

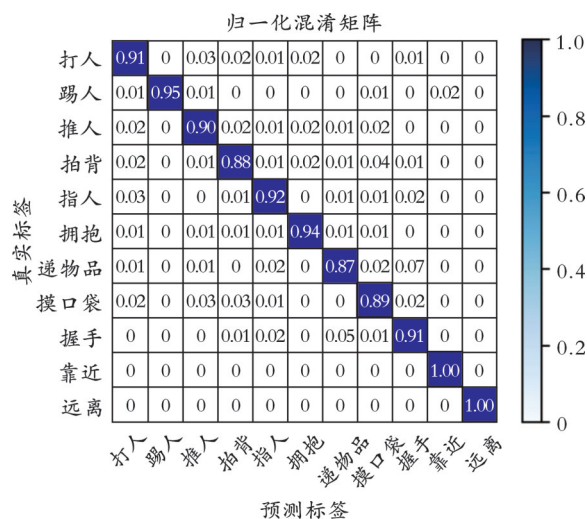


图6 混淆矩阵

从图6可以看出,“拍背”、“递物品”、“摸口袋”这三类双人交互动作的识别率较低,因为这三类动作主要集中在一个人身上,另外一个人的动作幅度较小。不难发现,“握手”和“递物品”两个动作容易混淆,主要是因为人体的骨架数据不包括递物品的有效表示,所以导致这两个动作较为相似,不易识别。

3.3 对比实验

为了验证建立双人骨架数据的交互依赖关系,验证本文所提算法的有效性,基于NTU RGB+D数据库,采用交叉验证(Cross-Validation,CV)的方式,进行了与近期研究成果的对比实验。在双人交互骨架数据上进行算法测试,与其他图卷积方法得到的行为识别结果进行了对比,实验结果如表2所示。

表 2 对比实验结果

方法	准确率/%
Clips+CNN+MTLN ^[5]	84.8
TS-LSTM ^[7]	81.2
IndRNN ^[8]	87.9
EleAtt-RNN ^[9]	87.1
ST-GCN ^[11]	88.3
DPRL+GCNN ^[13]	89.8
motif-GCNs+non-local VTDB ^[12]	90.2
DHI-STGCN	92.4

由表 2 可知,本方法在 NTU 数据集的 CV 评价标准上的识别准确率为 92.4%,相比于 CNN 和 RNN 的网络,图卷积通过附近其他节点的信息得到该节点的信息,本质上是传播特征,而不是传播标签;通过已知的特征节点得到该节点的属性,既能

学习节点特征,又能自动地学习节点之间的重要关联信息。相比于文献[9]、[11]、[12]和[13]的图卷积方法,本文构建的骨架图充分考虑了双人的交互信息,且 SRGCN 网络中的图注意机制不仅提取了更优的空间特征,还增加了对关键帧的关注,得到了更高的识别准确率。

4 结 论

提出了一种双人交互时空图卷积神经网络算法(DHI-STGCN),融合了时间信息与空间结构信息。在图的创建过程中,使用双人骨架数据源,构造了空间动作图,充分考虑了双人交互的依赖关系,在时间信息上构建了时序信息图,反映了交互动作行为的动态变化过程,提高了其时序建模能力。实验结果表明,该方法可以有效地考虑两个人之间的交互信息,在识别准确率上有一定的提升。

参 考 文 献

- [1] 张宇,曹江涛,阚哲,等.基于CPS系统的网络入侵检测方法[J].传感器与微系统,2020,39(10):126-128.
- [2] 魏鹏,曹江涛,姬晓飞.基于整体和个体分割融合的双人交互行为识别[J].辽宁石油化工大学学报,2019,39(6):91-96.
- [3] 姬晓飞,王治博,王昱.视频手势识别的互动演示系统的设计与实现[J].沈阳航空航天大学学报,2020,37(2):56-62.
- [4] 陈姝琪,曹江涛,赵挺,等.基于关节点数据的双人交互行为识别[J].电子测量与仪器学报,2020,34(6):124-130.
- [5] Ke Q, Bennamoun M, An S, et al. A new representation of skeleton sequences for 3D action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Amsterdam:[s.n.], 2017.
- [6] Liu J, Shahroudy A, Xu D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]//Proceeding of the European Conference on Computer Vision. Amsterdam:[s.n.], 2016.
- [7] Lee I, Kim D, Kang S, et al. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice:[s.n.], 2017.
- [8] Li S, Li W, Cook C, et al. Independently recurrent neural network (INDRNN): Building a longer and deeper RNN[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:[s.n.], 2018.
- [9] Zhang P F, Xue J R, Lan C L, et al. Eleatt-RNN: Adding attentiveness to neurons in recurrent neural networks[C]//European Conference on Computer Vision. Munich:[s.n.], 2018.
- [10] Wang P, Li Z, Hou Y, et al. Action recognition based on joint trajectory maps using convolutional neural networks[C]//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam:[s.n.], 2016.
- [11] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans:[s.n.], 2018.
- [12] Wen Y H, Gao L, Fu H, et al. Graph CNNs with motif and variable temporal block for skeleton-based action recognition [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii:[s.n.], 2019.
- [13] Tang Y, Yi T, Lu J, et al. Deep progressive reinforcement learning for skeleton-based action recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City:[s.n.], 2018.
- [14] Shahroudy A, Liu J, Ng T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]//IEEE Computer Society.[s.l.]:IEEE, 2016.

(编辑 张 芳)