智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

交互关系超图卷积模型的双人交互行为识别

代金利, 曹江涛, 姬晓飞

引用本文:
代金利,曹江涛,姬晓飞. 交互关系超图卷积模型的双人交互行为识别[J]. 智能系统学报, 2024, 19(2): 316–324.
DAI Jinli, CAO Jiangtao, JI Xiaofei. Two-person interaction recognition based on the interactive relationship hypergraph convolution network model[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 316–324.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202208001

DOI: 10.11992/tis.202208001
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20231110.1512.006

交互关系超图卷积模型的双人交互行为识别

代金利[1], 曹江涛[1], 姬晓飞[2]
（1. 辽宁石油化工大学 信息与控制学院, 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: 为提高学校、商场等公共场所的安全性, 实现对监控视频中的偷窃、抢劫和打架斗殴等异常双人交互行为的自动识别, 针对现有基于关节点数据的行为识别方法在图的创建中忽略了 2 个人之间的交互信息, 且忽略了单人非自然连接关节点间的交互关系的问题, 提出一种基于交互关系超图卷积模型用于双人交互行为的建模与识别。首先针对每一帧的关节点数据构建对应的单人超图以及双人交互关系图, 其中超图同时使多个非自然连接节点信息互通, 交互关系图强调节点间交互强度。将以上构建的图模型送入时空图卷积对空间和时间信息分别建模, 最后通过 SoftMax 分类器得到识别结果。该算法框架的优势是在图的构建过程中加强考虑双人的交互关系、非自然连接点间结构关系以及四肢灵活的运动特征。在 NTU 数据集上的测试表明, 该算法得到了 97.36% 的正确识别率, 该网络模型提高了双人交互行为特征的表征能力, 取得了比现有模型更好的识别效果。

关键词: 双人交互; 行为识别; 关节点数据; 深度学习; 时空图卷积网络; 超图; 图结构; 神经网络
中图分类号: TP183　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)02−0316−09

中文引用格式: 代金利, 曹江涛, 姬晓飞. 交互关系超图卷积模型的双人交互行为识别 [J]. 智能系统学报, 2024, 19(2): 316–324.
英文引用格式: DAI Jinli, CAO Jiangtao, JI Xiaofei. Two-person interaction recognition based on the interactive relationship hypergraph convolution network model[J]. CAAI transactions on intelligent systems, 2024, 19(2): 316–324.

Two-person interaction recognition based on the interactive relationship hypergraph convolution network model

DAI Jinli[1], CAO Jiangtao[1], JI Xiaofei[2]
(1. School of Information and Control Engineering, Liaoning Petrochemical University, Fushun 113001, China; 2. School of Automation, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: To enhance the security of schools, shopping malls, and other public places, it is important to achieve automatic identification of abnormal two-person interaction behaviors, such as stealing, robbing, fighting, and assaulting, in surveillance videos. However, the current behavior recognition method based on joint data in graph creation neglects the two-person interaction information as well as the interaction relationship between the single unnatural connection joints. To address this issue, a two-person interaction behavior recognition model based on the interactive relationship hypergraph convolution network is proposed to model and identify human interactions. First, the corresponding single hypergraph and two-person interaction graph are created for the joint-point data of each frame, where the hypergraph makes the information of multiple unnaturally connected nodes interchangeable at the same time, and the interaction graph emphasizes the interaction strength between nodes. The above-constructed graph models are fed into the spatiotemporal graph convolution to model the spatial and temporal information separately, and lastly, the recognition results are acquired by the SoftMax classifier. The benefits of the proposed algorithm framework are that the interactive relationship between two persons, the structural relationship between unnatural connections, and the flexible motion characteristics of limbs are regarded in the graph construction process. Tests on the NTU data set demonstrate that the algorithm attains a correct recognition rate of 97.36%. The findings indicate that the network model enhances the ability to represent the characteristics of two-person interaction and has better recognition performance than the current models.

Keywords: two-person interaction; behavior recognition; skeleton node data; deep learning; ST-GCN; hypergraph; graph structure; neural networks

随着计算机科学技术的迅速发展, 基于视频的双人交互行为识别已经成为计算机视觉领域的研究热点[1-2], 且取得了一定的研究进展。RGB

收稿日期: 2022−08−01.　**网络出版日期:** 2023−11−13.
基金项目: 国家自然科学基金项目 (61673199); 辽宁省科技公益研究基金项目 (2016002006).
通信作者: 姬晓飞. E-mail: jixiaofei7804@126.com.

视频获取简单且包含丰富的外观信息, 但缺少深度维度信息, 对于复杂行为的识别准确性不高[3]。Kinect 设备获取的人体三维关节点数据表达信息全面, 且数据简便, 不仅可以记录每个人关节的运动信息, 也可以记录双人关节之间的交互信息[4-6]。深度学习的方法与关节点数据有效融合, 为提高双人交互行为识别的精度提供了新的解决方案。

目前针对骨架关节点数据, 基于深度学习识别方法主要分为三类: 基于循环神经网络 (recurrent neural network, RNN) 的方法[7]、基于卷积神经网络 (convolutional neural networks, CNN) 的方法[8]、基于图卷积神经网络 (graph convolution network, GCN) 的方法[9]。其中基于 RNN 的方法侧重对行为时序关系的表示, Liu 等[10] 提出时空长短时记忆网络 (spatial-temporal long short-term memory, ST-LSTM), 通过对原始关节数据进行树状结构排列建立关节点间关系, 送入 LSTM 进行时序建模与识别, 该类方法往往缺少空间信息的合理表示。基于 CNN 的方法试图学习双人之间交互的动态表征, Choutas 等[11] 将关节点序列以颜色编码的方式进行图像化, 并用 CNN 对图像提取特征, 该算法实现简单, 但是其以颜色编码方式压缩时序维度, 造成了无法弥补的信息损失。

基于 CNN 和 RNN 的方法往往将原始骨架数据转化为网格状的输入, 只考虑卷积核内相邻的共现特征, 不能充分利用关节点数据的结构信息。近几年, 基于 GCN 的方法以一种更灵活的方式处理关节点数据, 将卷积从图像推广到图, 可以很好地探索关节点之间的结构关系。Yan 等[12] 首先将 GCN 引入基于关节点的动作识别中, 提出了时空图卷积网络 (spatial temporal GCN, ST-GCN), 构造一个以关节点为顶点, 以人体结构自然连接为边的时空图, 用 SoftMax 分类器将图中高级特征映射为相应的动作类别。在此基础上, Cheng 等[13] 提出了一种改进的移位图进化网络 (shift-GCN) 来增强空间图的表达能力, 其中移位图操作为空间图和时间图提供了灵活的感受野。研究表明行为识别中动作在空间域上的变化幅度要大于在时间域上的, 因此 Song 等[14] 提出丰富的激活图卷积网络 (richly activated GCN, RAGCN), 该网络通过对每个邻接矩阵学习新的权重来突出边的重要性。以上方法在双人行为识别中独立识别单人构造的自然连接图, 未考虑非自然连接点 (如手、脚) 交互关系和双人之间交互关系, 在交互动作识别中没有显式地构建图结构, 难以学习无物理联系的关节间的关系[15]。针对 ST-GCN 预定图中非自然连接, 距离远且信息不互通的问题, Li 等[16] 提出动作结构图卷积网络 (actional-structural GCN, AS-GCN), 通过引入额外的邻接矩阵的方式, 建立动作相关的依赖, 这种表征动作连接的邻接矩阵可以视为一种新的、用来表征非自然连接关节点间动作相关性的邻接矩阵。为了表示双人交互关节点间的关系, 成科杨等[17] 把两人作为一个整体, 将两人间交互关节连接为边构建为交互连接图, 保留了双人运动的局部交互信息。Wu 等[18] 提出行动者关系图卷积 (actor relation GCN, ARGCN) 对关系图进行关系推理, 根据个体的位置和特征, 以个体之间的连接为边建立多个关系图。将各个关系图的结果融合在一起, 生成所有参与者个体的关系表示, 分别进行个体行为识别和群组行为识别。Chen 等[19] 提出通道拓扑细化图卷积网络 (channel-wise topology refinement GCN, CTR-GCN) 来动态学习不同的拓扑图, 并有效地聚合不同通道的关节特征。Lee 等[20] 提出层次分解图卷积网络 (hierarchically decomposed GCN, HD-GCN) 将每个关节节点分解为多个集合, 提取主要的相邻边和远边, 并利用其构建包含这些边的 HD-Graph, 令这些边位于人体骨骼的相同语义空间中。

根据以上分析, 为解决预定的关节点连接图仅表示人体物理结构, 对非自然连接和交互关节点间的关系无法有效表达的问题, 现阶段研究试图直接连接单人非自然连接点和双人间交互关节点作为普通边, 来获取非自然连接点间和交互关节点间信息, 但该方法只能表达关节点间物理结构关系, 未能有效利用灵活的四肢关系特征以及双人交互相关性特征。

为有效利用关节点运动特征识别双人交互动作, 提出一种基于交互关系超图卷积模型的双人交互行为识别算法。首先针对仅用普通边连接成的图结构无法有效表达非自然连接点间信息的问题, 构造以四肢为超边的人体图结构, 有效利用最能区分动作类别的四肢运动特征, 且用超图结构打破传统图结构的局限性, 实现多个非自然关节点间信息互通。再针对用普通边连接成的图结构无法有效表达双人交互信息的问题, 构造双人交互关系图结构, 通过提取交互关节之间连接的强度关系作为交互动作特征。最后将构建的人体超图和交互关系图嵌入 ST-GCN, 将图结构与时空图卷积相结合, 更有效地提取关节点间以及帧间的依赖信息。试验结果表明, 所提算法在识别

率上得到大幅度的提升，同时与现有算法比较，验证了所提方法的有效性。

1　算法框架

本研究提出一种基于交互关系超图卷积模型的双人交互行为识别算法，该算法将以四肢为超边的人体超图和双人交互关系图与 ST-GCN 网络结合，算法框架如图 1 所示。

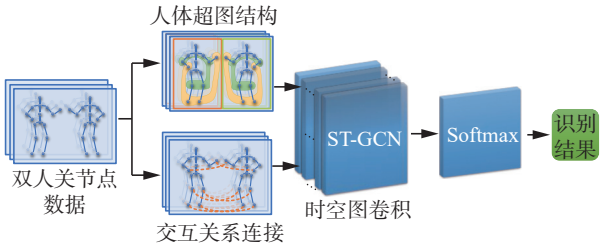


图 1　算法整体流程

Fig. 1　Overall flow of proposed algorithm

主要步骤包括：

1) 构建人体超图。首先对人体四肢构建超边，将四肢超边与人体自然连接图联合构建出人体超图，用于表示人体非自然连接点之间交互关系，充分利用运动时最灵活的四肢关节点特征，并用超图概念实现关节点间信息互通。

2) 构建交互关系图。构建双人交互连接，通过计算双人各个对应关节点的反向距离，进而推算出交互的连接强度矩阵，充分考虑突出双人间交互行为信息。

3) 人体超图和交互关系图嵌入 ST-GCN。将构建的人体超图与交互关系图嵌入 ST- GCN，利用 ST-GCN 对所创建的交互身体模型提取四肢交互特征和双人交互行为的时空特征。

4) 识别分类。将 ST-GCN 处理后得到的特征向量送入 SoftMax，生成分类概率进行交互识别。

2　构建人体超图

现阶段研究中使用的关节点连接图是预定义的，仅表示人体物理结构，对非自然连接点间的交互信息无法有效表达。例如走路，双手双脚之间的关系很重要，但是 ST-GCN 预定义的人体图中彼此距离很远，而且图中一条边只能包含 2 个关节点，因此不能在四肢间实现信息互通。这就是传统图结构的局限性，其很容易忽视掉一些高阶结构信息。为了尽可能地保存高阶信息，引入超图 (Hypergraph) 这一工具[21-22]。超图是一种广义上的图，其边可以和任意数量的顶点连接。

定义人体超图由自然连接的关节点和构建的四肢超边构成，如图 2(a) 所示。其中四肢超边用 $G_H=(V,E_h)$表示，V代表一帧中所有的关节点构成的集合，E_h是V的非空子集称为超边集合，使用超边连接多个相关的关节，包括非自然连接的关节，如手和脚，如图 2(b) 所示。Yadati 等[23]提出了具有相同潜在动机的超图拉普拉斯，其讨论的超图的一个特殊方面是，每个超边 e 都是由单个的一对简单边$\{i_e,j_e\}$表示的，这个简单边可能会随着时间的变化而变化。已有研究表明，广义超图拉普拉斯算子满足文献 [23] 给出的上述拉普拉斯算子所满足的所有性质，借鉴以上思想使用不带权值的初始特征构造带中间介质的超图拉普拉斯矩阵，具有中间介质的超图拉普拉斯中每个超边中简单边的权值之和为 1，如图 2(c) 所示。

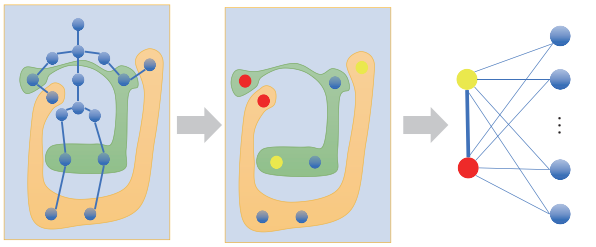


图 2　单人超图结构的创建

Fig. 2　Creation of single-player hypergraph structures

常用的双人交互行为识别数据集中关节点标号如图 3 所示，四肢作为超边的人体超图计算过程如下：首先定义四肢超边的关系矩阵为 $\boldsymbol{A}_3=\mathrm{diag}(\boldsymbol{A}_1,\boldsymbol{A}_2)$，其中$\boldsymbol{A}_1$和$\boldsymbol{A}_2$分别表示 2 个超边 $\boldsymbol{A}_1=\boldsymbol{A}_2=\begin{bmatrix}1&1&1&1\end{bmatrix}^{\mathrm{T}}$，$\boldsymbol{A}_1$中关节点序号分别为 8、12、15、9，$\boldsymbol{A}_2$中关节点序号分别为 6、10、14、18。定义超边度矩阵 $\boldsymbol{D}_e=\mathrm{diag}(4,4)$，矩阵大小为超边的个数，对角数值为每条超边中包含的节点个数，构建的人体超图结构包括 2 个超边，每条超边中有 4 个关节点。然后将各个超边权值分别设为ω_1，ω_2，则超边权值矩阵$\boldsymbol{W}=\mathrm{diag}(\omega_1,\omega_2)$。由此计算得到该矩阵的拉普拉斯矩阵$\boldsymbol{L}_1$。在人体自然连接普通图邻接矩阵基础上，赋予超图拉普拉斯矩阵权重，得到以四肢为超边的人体超图拉普拉斯矩阵$\boldsymbol{A}_H$。该超图的拉普拉斯矩阵依然是以节点与节点之间的关系为核心，但相较于普通图，不仅实现了非自然连接点间信息互通，而且每条边不局限于两点之间邻接关系，实现四肢权重互通。

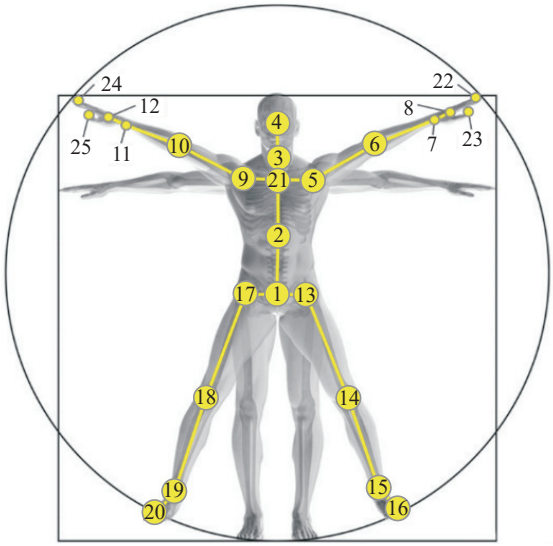
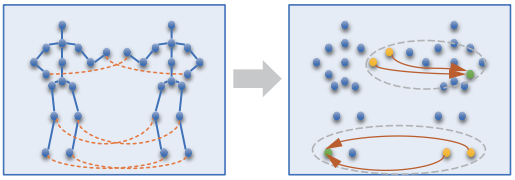


图 3　NTU 数据集人体关节点
Fig. 3　NTU data set human node

3　构建交互关系图

为充分考虑交互依赖关系，突出表示双人之间交互信息的重要性，双人交互特征的提取是对交互身体关节对的帧内特征的提取，通过计算第 1 个人的每个节点到第 2 个人的每个节点的反向距离，决定关系连接的强度，这些连接将分离骨架图中的节点连接起来，如图 4 所示，从而来捕捉两人之间的交互特征。



(a) 关节点交互连接　　(b) 交互关系运算

图 4　双人交互关系结构的创建
Fig. 4　The creation of a two-person interactive relationship structure

用$\hat{A}$表示几何关节相关性，计算公式为

$$\hat{A}[t,i,j]=\exp\left(--\frac{\left\|F_{\text{in}}\left(v_{0,i}^{t}\right)-F_{\text{in}}\left(v_{1,j}^{t}\right)\right\|^{2}}{C_{\text{in}}}\right) \tag{1}$$

式中：$v_{0,i}^{t}$、$v_{1,j}^{t}$分别表示第 t 帧中两人的第 i 个关节；C_{in}为输入通道数；$F_{\text{in}}\left(v_{0,i}^{t}\right)$、$F_{\text{in}}\left(v_{1,j}^{t}\right)$分别为 2 个人的输入特征映射，定义$F_{\text{in}}\left(v_{m,i}^{t}\right)=S_{m,i}^{t}\left(S_{m,t}^{t}\in \mathrm{R}^{M\times C_m\times T\times N}\right)$。$F_{\text{in}}\left(v_{m,i}^{t}\right)$测量了两人之间各个节点的反向距离，$v_{0,i}^{t}$、$v_{1,j}^{t}$之间的欧氏距离越小，$\hat{A}$值越大，代表交互关系越强。

为了防止过度拟合和过滤掉不相关的连接，用整流单元直接过滤掉不相关的连接。这个操作有效地消除薄弱环节，从而强调最相关的环节，只保留大多数相关连接，可以帮助模型提取出最具识别性的交互特征。比如，握手时双方脚没动，只保留上肢交互特征。整流公式被定义为

$$\overline{A}[t,i,j]=\begin{cases}\hat{A}[t,i,j]\ ,\ \hat{A}[t,i,j]\geqslant 0.5\\ 0,\ \text{其他}\end{cases} \tag{2}$$

最后用方程归一化为

$$\boldsymbol{A}_{I}=\boldsymbol{D}_{\text{r}}^{-\frac{1}{2}}\bar{\boldsymbol{A}}\boldsymbol{D}_{\text{c}}^{-\frac{1}{2}} \tag{3}$$

得出的双人交互关系矩阵$\boldsymbol{A}_I$，用来在数学上表示动作序列中所有帧的关系图和人体超图关系图。

4　人体超图和交互关系图嵌入 ST-GCN

ST-GCN 网络结构如图 5 所示，首先做归一化 (batch normalization, BN) 处理，并在进行图卷积之前，加入注意力模型 (attention model, AM)。在运动过程中，不同的关节点重要性是不同的，因此，ST-GCN 对不同关节点进行了加权，且每个 ST-GCN 单元都训练不同的权重参数。接着，该模型通过不断堆叠 ST-GCN，从图结构输入中持续提取高级的语义特征，交替使用 GCN 和时间卷积网络 (temporal convolutional network, TCN)，对时间和空间维度进行变换。最后，引入全局平均池化 (global average pooling, GAP) 以及全连接层 (fully convolutional network, FCN) 输出预测分支。

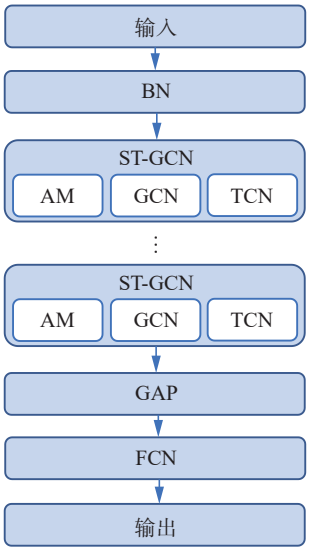


图 5　ST-GCN 网络结构
Fig. 5　ST-GCN network structure diagram

嵌入时空图卷积融合人体超图结构和双人交互关系图结构，以此捕捉关节点之间更丰富的相

关性特征，并提取到更丰富有效的判别性特征。图结构嵌入时空图卷积的具体结构如图 6 所示，其结构由空间卷积和时间卷积交替组合而成。为了提取到丰富的关节点相关特征，空间卷积操作由 2 个分支组成，分别是人体超图、双人交互关系图所对应的关系矩阵$\boldsymbol{A}_H$、$\boldsymbol{A}_I$。

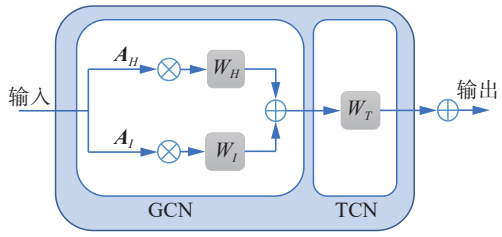


图 6　图结构嵌入时空图卷积结构

Fig. 6　Graph structure embedded space-time graph convolution

空间卷积操作分别用图卷积提取 2 个图结构的特征，然后对其特征通过对应元素逐点相加的方式进行融合。用图集$G=\{H,I\}$表示其组合，空间卷积操作可以被表示为

$$\boldsymbol{F}_{\text{out}}=\sum_{g\in G}\sigma(\boldsymbol{W}_g\boldsymbol{F}_{\text{in}}(\boldsymbol{A}_g\circ\boldsymbol{M}_g)) \tag{4}$$

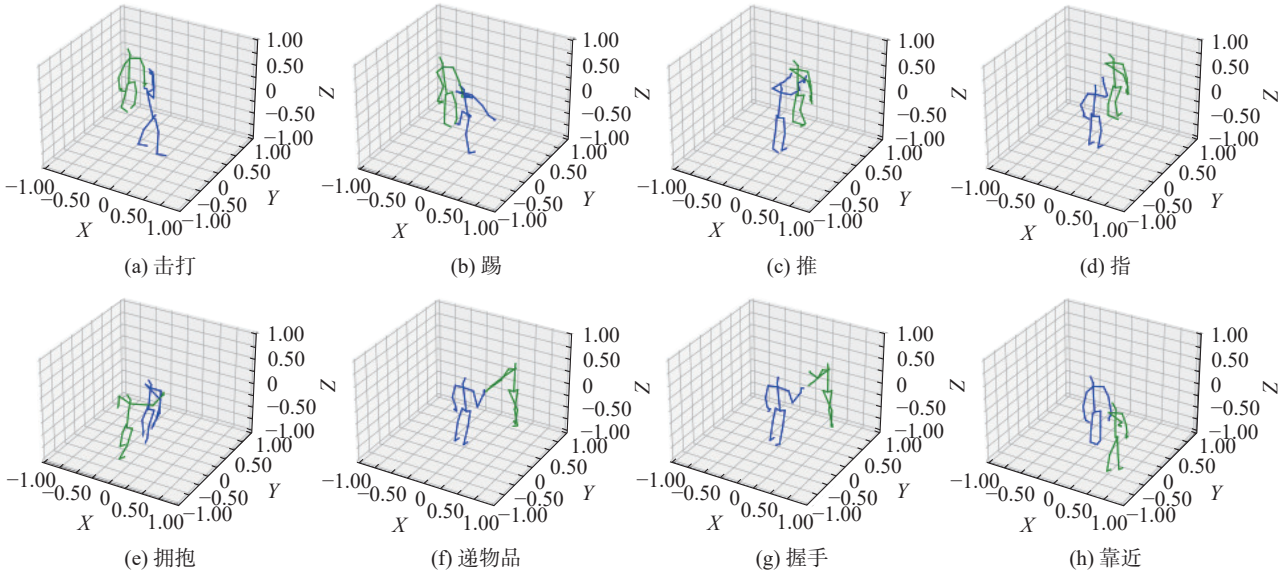
式中：$\boldsymbol{A}_g$ 代表图邻接矩阵；$\boldsymbol{M}_g$代表可学的权重分配重要性；$\boldsymbol{F}_{\text{in}}$是输入特征向量；$\boldsymbol{W}_g$代表边缘惊醒线性特征变换；$\circ$代表逐元素相乘；$\sigma$是修正线性单元 (rectified linear unit, ReLu)；$\boldsymbol{F}_{\text{out}}$是输出特征向量。

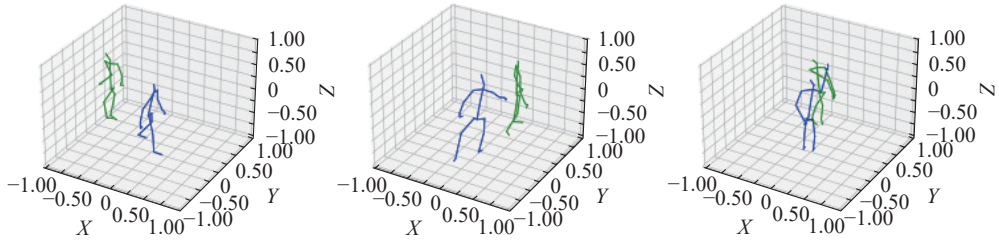
将人体超图和交互关系图嵌入至 ST-GCN 单元中，矩阵在 GCN 中进行运算，实现空间维度信息的聚合，利用 TCN 网络实现时间维度信息的聚合。叠加 10 个 ST-GCN 单元，这些 ST-GCN 单元具有不同的输出通道和时间跨度。前 4 个 ST-GCN 单元有 64 个输出通道，中间 3 个有 128 个输出通道，后 3 个有 256 个通道。第 5 和第 8 个 ST-GCN 单元的时间跨度是 2，其他的是 1。实现双人交互关系图结构和超图结构与时空图卷积结合成为交互关系超图卷积模型。

5　试验结果分析

5.1　数据集介绍

在 NTU RGB+D[24] 数据集上进行训练与测试，该数据集是目前最大的行为识别数据集，是利用 Kinect v2 相机获得，包含 RGB 视频帧、深度信息和 3D 关节点信息。该数据集包含 56 000 个视频序列，共 60 类动作。包括 40 类日常行为、9 类医疗健康相关行为以及 11 类双人交互行为。采用关节点数据中的 11 类交互动作 (Mutual)，对所提算法进行评估，即击打、踢、推、指、拥抱、递物品、握手、靠近、远离、摸口袋和拍背，骨架数据结构图如图 7 所示。该数据集的官方评估方法的协议有 2 种类型：交叉受试者（cross-subject, CS）和交叉视角（cross-view, CV）。CV 按相机编号划分，三台相机角度分别为−45°、0°、45°，为应对实际场景中视角不同且多变的情形，本试验根据 CV 协议进行评估，共采用 8 408 个双人交互行为视频数据，训练集是由 5 606 个 2 号摄像头和 3 号摄像头捕获的视频，验证集是由 2 802 个 1 号摄像头捕获的视频，训练集与验证集比例为 2∶1。





(i) 远离　　(j) 摸口袋　　(k) 拍背

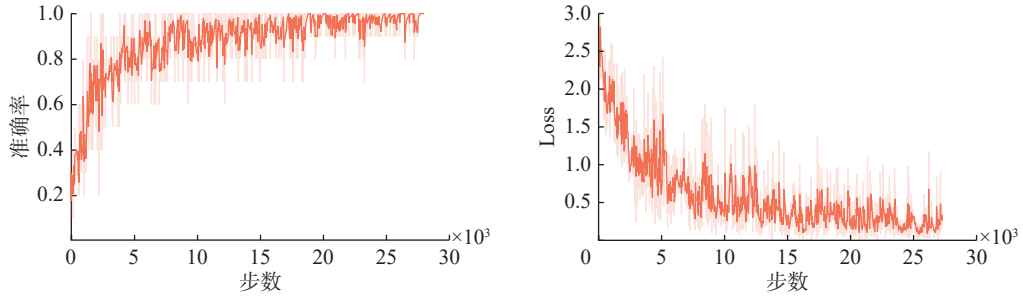
图 7　NTU RGB+D 骨架数据示例

Fig. 7　NTU RGB+D skeleton data example

5.2　试验结果

遵循文献 [25] 中 NTU-RGB+D 的数据预处理过程。输入张量的形状为 M (2)×C (3)×T (300)×N (25)。对于一个 batch 的视频用 4 维矩阵 (M, C, T, N) 表示, 其中 M 代表视频中实施动作的人数, C 代表关节的特征维度, T 代表一个视频帧的数量, N 代表关节的数量, 这里是 25 个关节。

试验在 Ubuntu16.04 操作系统下进行, 采用基于 Python3.7 的深度学习框架 Pytorch, GPU 为 NIVIDIA 1080 Ti 的深度学习环境。将 NTU 数据集中交互部分的关节点数据以行为类别进行划分, 按照 CV 标准进行训练测试。经过时空卷积得到的张量用 BN 层归一化, 并由 ReLu 层激活。由于硬件的限制, 对 NTU-RGB+D (Mutual) 的批处理大小设定为 10, 初始学习率设置为 0.01, 使用交叉熵损失函数计算损失, 并使用随机梯度下降法 (stochastic gradient descent, SGD) 算法进行优化。本试验采用 50 次迭代进行训练, 得到的训练集和测试集对应的准确率和损失函数的变化曲线如图 8 所示。



(a) 准确率变化曲线　　(b) 损失函数值变化曲线

图 8　交互关系超图卷积模型训练准确率和损失函数值

Fig. 8　Interactive relationship hypergraph convolution model training accuracy and loss function value

由图 8 可以看出, 模型最终收敛趋于稳定, 最终得到的最高识别率为 97.36%。损失函数值为 0.09188。为了进一步分析模型的性能, 进行了对比试验。

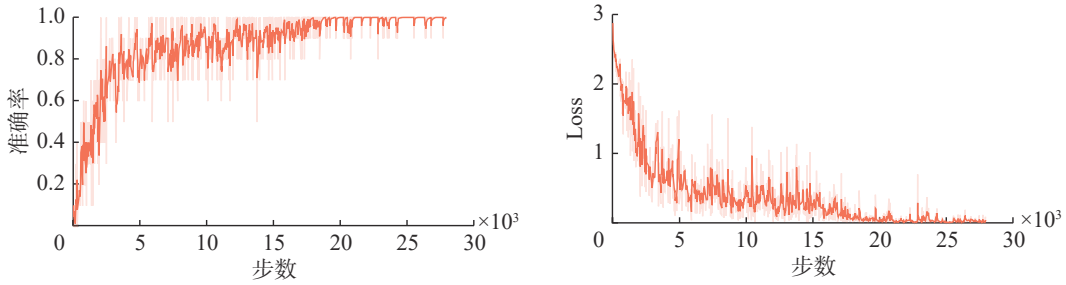
5.3　对比分析

为了进一步验证该算法的有效性, 将分 2 种情况对试验结果进行对比分析。首先验证双人交互连接强度的有效性, 其次验证人体超图的有效性, 以上试验在 NTU RGB+D (Mutual) 数据库下进行训练与测试。

5.3.1　验证双人交互连接强度的有效性

ST-GCN 模型在 NTU-RGB+D(Mutual) 的 CV 基准上进行试验, 加入双人交互关系矩阵后在同一环境下试验, 对应的准确率和损失函数的变化曲线如图 9 所示, 生成混淆矩阵分别如图 10 图 11 所示。

从上面 ST-GCN 的混淆矩阵可以看出, 在指、拍、推、打这几类上肢动作之间, 如果不提取交互关系的情况下会出现大量混淆。这是由于 ST-GCN 单独提取每个人体特征, 而不是同时提取两人的交互特征, 在指、拍背、推、打这几类动作, 都是一人不动, 另一人抬起上肢, 对于单独个体来说动作相似, 因此容易混淆。加入双人交互关系后, 以上几类容易混淆的交互动作分类出错情况明显减少, 这些交互特性有助于更好地区分交互行为。证明了双人交互关系图有效地提取了交互信息特征, 且交互关系图与 ST-GCN 模型是高度兼容的。



(a) 准确率变化曲线　(b) 损失函数值变化曲线

图 9　交互关系图卷积模型训练准确率和损失函数值

Fig. 9　Interactive relationship graph convolution model training accuracy and loss function value

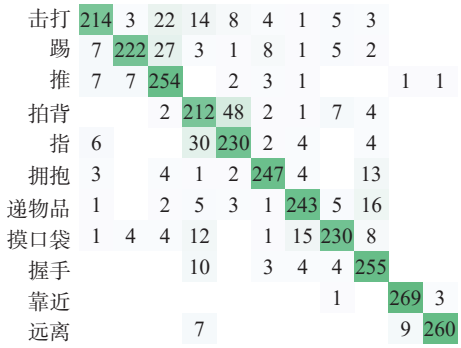


图 10　ST-GCN 混淆矩阵

Fig. 10　ST-GCN confusion matrix

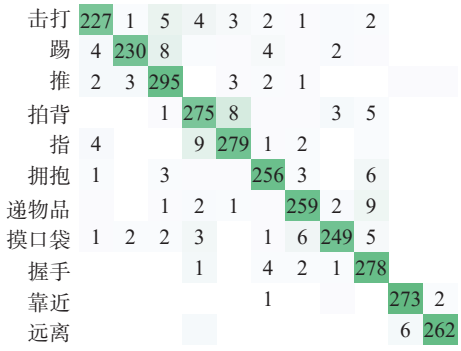
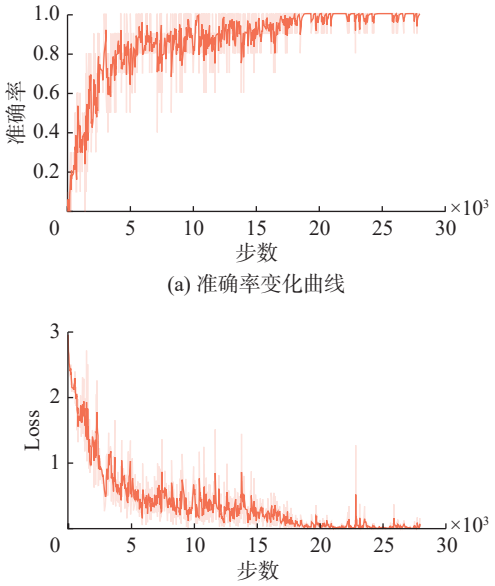


图 11　引入交互关系混淆矩阵

Fig. 11　ST-GCN(Interaction) confusion matrix

5.3.2　验证人体超图的有效性

在加入双人关系的基础上，加入人体超图结构，观察其对交互行为识别的贡献，对应的准确率和损失函数的变化曲线如图 12 所示。由图 13 的 3 个模型在 11 种交互动作上的准确率对比可以看出，加入交互连接强度关系后，有相似特性的上肢交互行为类别准确率得到大幅度提升，加入超图后的模型在每类交互动作识别中都有些提升，尤其在击打、踢、拥抱这些四肢运动幅度大的动作识别更准确，说明超图更有效表达个体运动特征，从而使交互行为识别准确率有所提升。



(a) 准确率变化曲线

(b) 损失函数值变化曲线

图 12　人体超图卷积模型训练准确率和损失函数值

Fig. 12　Body hypergraph convolution model training accuracy and loss function value

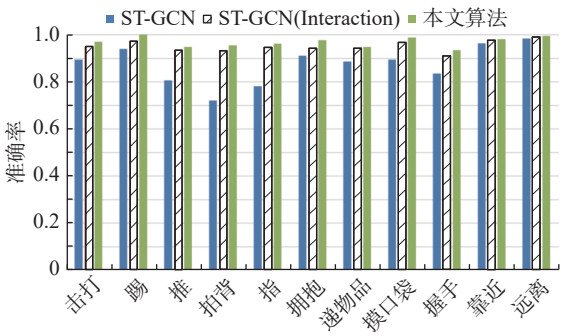


图 13　比较 3 个模型在 11 种交互动作上的准确率

Fig. 13　Accuracy of the three models on 11 kinds of interaction actions was compared

比较这 3 种算法在 NTU-RGB+D (Mutual) CV 上的分数，如表 1 所示，与基线模型 ST-GCN 相比，加入双人交互图结构后识别准确率提高了 4.61 个百分点，加入交互图结构与超图结构后准确率比 ST-GCN 提高了 6.34 个百分点。试验结果证明了人体超图结构和双人交互图结构的有效性。

表 1　消融结果对比
Table 1　Comparison of ablation results

模型名称	双人交互	超图结构	识别结果/%
ST- GCN	×	×	91.02
ST-GCN(Hypergraph)	×	√	94.69
ST-GCN(Interaction)	√	×	95.63
本研究算法	√	√	**97.36**

5.4　与其他算法对比

为了验证所提出模型的有效性, 将试验结果与基于关节点数据的其他方法在 NTU RGB+D 数据库下进行试验的结果对比, 如表 2 所示。

表 2　本研究模型与其他模型算法结果对比
Table 2　The results of this model are compared with those of other models

识别方法	识别结果/%
ST-LSTM[10]	87.30
ST-GCN[12]	91.02
AS-GCN[16]	93.46
3S RA-GCN[14]	93.60
4S Shift-GCN[13]	96.50
本研究算法	**97.36**

由表 2 可以看出, 用 GCN 处理关节点数据比文献 [10] 采用 ST-LSTM 的方法准确率有了大幅度的提高, 说明加强对图结构的学习可以更有效地提取基于关节点的行为特征。文献 [12-14, 16] 提出的双人交互行为识别算法均采用 ST-GCN 及改进的图结构算法, 但未考虑多个非自然连接关节点间的信息交互问题以及双人之间的交互关系, 导致准确率并没有明显突破。与以上算法相比, 本研究提出的基于交互关系超图卷积模型的双人交互行为识别算法通过创新图结构实现多个非自然连接节点间信息交互, 获得了最好的识别结果, 验证了此模型的优越性。

6　结束语

本研究提出基于交互关系超图卷积模型的双人交互行为识别算法, 设计交互关系矩阵来表示交互关系图结构, 结合几何特征表达 2 个人关节点的交互强度关系。设计以四肢为超边的超图拉普拉斯矩阵来表示人体超图结构, 实现多个非自然连接关节间交互信息, 强调了运动时的人体协调特征。通过多层时空卷积层来构建网络的主干, 充分利用运动时人体关节间的空间和时间依赖性, 发现关节点之间的潜在关系, 从而更高效识别, 同时证明了上述图结构与 ST-GCN 模型的兼容性。试验证明所提出的模型在双人交互识别方面表现出卓越的能力, 在 NTU 的交互数据集上得到了令人满意的识别精度。

参考文献:

[1] 吴联世, 夏利民, 罗大庸. 人的交互行为识别与理解研究综述 [J]. 计算机应用与软件, 2011, 28(11): 60–63.
WU Lianshi, XIA Limin, LUO Dayong. Survey on human interactive behaviour recognition and comprehension[J]. Computer applications and software, 2011, 28(11): 60–63.

[2] WANG Pichao, LI Wanqing, OGUNBONA P, et al. RGB-D-based human motion recognition with deep learning: a survey[J]. Computer vision and image understanding, 2018, 171: 118–139.

[3] BARADEL F, WOLF C, MILLE J, et al. Glimpse clouds: human activity recognition from unstructured feature points[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 469−478.

[4] YUN K, HONORIO J, CHATTOPADHYAY D, et al. Two-person interaction detection using body-pose features and multiple instance learning[C]//2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence: IEEE, 2012: 28−35.

[5] 姬晓飞, 谢旋, 任艳. 深度学习的双人交互行为识别与预测算法研究 [J]. 智能系统学报, 2020, 15(3): 484–490.
JI Xiaofei, XIE Xuan, REN Yan. Research on two-person interaction Recognition and Prediction Algorithm based on Deep Learning[J]. CAAI transactions on intelligent systems, 2020, 15(3): 484–490.

[6] HUYNH-THE T, BANOS O, LE B V, et al. PAM-based flexible generative topic model for 3D interactive activity recognition[C]//2015 International Conference on Advanced Technologies for Communications. Ho Chi Minh City: IEEE, 2016: 117−122.

[7] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[EB/OL]. (2014−09−18)[2022−01−01]. https://arxiv.org/abs/1409.2329.pdf.

[8] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[9] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL].(2017−09−09)[2022−01−01]. https://arxiv.org/abs/1607.07043.pdf.

[10] LIU Jun, SHAHROUDY A, XU Dong, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[EB/OL]. (2016−07−24)[2022−01−01]. https://arx-

iv.org/abs/1607.07043.pdf.

[11] CHOUTAS V, WEINZAEPFEL P, REVAUD J, et al. PoTion: pose MoTion representation for action recognition[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7024−7033.

[12] YAN Sijie, XIONG Yuanjun, LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[EB/OL]. (2018−01−25)[2022−01−01]. https://arxiv.org/abs/1801.07455.

[13] CHENG Ke, ZHANG Yifan, HE Xiangyu, et al. Skeleton-based action recognition with shift graph convolutional network[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 180−189.

[14] SONG Yifan, ZHANG Zhang, SHAN Caifeng, et al. Richly activated graph convolutional network for robust skeleton-based action recognition[J]. IEEE transactions on circuits and systems for video technology, 2021, 31(5): 1915–1925.

[15] 刘云, 薛盼盼, 李辉, 等. 基于深度学习的关节点行为识别综述 [J]. 电子与信息学报, 2021, 43(6): 1789–1802.
LIU Yun, XUE Panpan, LI Hui, et al. A review of action recognition using joints based on deep learning[J]. Journal of electronics & information technology, 2021, 43(6): 1789–1802.

[16] LI Maosen, CHEN Siheng, CHEN Xu, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 3590−3598.

[17] 成科扬, 吴金霞, 王文杉, 等. 融合时空图卷积的多人交互行为识别 [J]. 中国图象图形学报, 2021, 26(7): 1681–1691.
CHENG Keyang, WU Jinxia, WANG Wenshan, et al. Multi-person interaction action recognition based on spatio-temporal graph convolution[J]. Journal of image and graphics, 2021, 26(7): 1681–1691.

[18] WU Jianchao, WANG Limin, WANG Li, et al. Learning actor relation graphs for group activity recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 9956−9966.

[19] CHEN Yuxin, ZHANG Ziqi, YUAN Chunfeng, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2022: 13339−13348.

[20] LEE J, LEE M, LEE D, et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition[EB/OL]. (2022−12−25)[2023−01−01]. https://arxiv.org/abs/2208.10741.pdf.

[21] ZHOU D, HUANG J, SCHÖLKOPF B. Learning with hypergraphs: clustering, classification, and embedding[C]//International on Neural Information Processing Systems. Vancouver MIT Press, 2006: 1601−1608.

[22] ZHOU Dengyong, HUANG Jiayuan, SCHÖLKOPF B. Learning from labeled and unlabeled data on a directed graph[C]//Proceedings of the 22nd International Conference on Machine Learning. New York: ACM, 2005: 1036−1043.

[23] YADATI N, NIMISHAKAVI M, YADAV P, et al. HyperGCN: a new method of training graph convolutional networks on hypergraphs[EB/OL]. (2018−09−07)[2022−01−01]. https://arxiv.org/abs/1809.02589.pdf.

[24] SHAHROUDY A, LIU Jun, NG T T, et al. NTU RGB D: a large scale dataset for 3D human activity analysis[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1010−1019.

[25] KE Qiuhong, AN Senjian, BENNAMOUN M, et al. SkeletonNet: mining deep part features for 3-D action recognition[J]. IEEE signal processing letters, 2017, 24(6): 731–735.

作者简介：

代金利，硕士研究生，主要研究方向为计算机视觉、图像处理和模式识别。E-mail：daijinli19980904@163.com。

曹江涛，教授，博士，主要研究方向为智能方法及其应用、视频分析与处理。主持国家自然科学基金项目 1 项、辽宁省自然科学基金项目 1 项。参与编著英文专著 2 部，发表学术论文 50 余篇。E-mail：jtcao@lnpu.edu.cn。

姬晓飞，副教授，博士，主要研究方向为视频分析与处理、模式识别理论。主持国家自然科学基金项目 1 项、辽宁省自然科学基金项目 1 项。参与编著英文专著 2 部，发表学术论文 40 余篇。E-mail：jixiaofei7804@126.com。