DOI:10.19651/j.cnki.emt.2210389

基于关节点运动估计的人体行为识别*

李志晗[1,2] 刘银华[2] 谢锐康[2] 单 良[2]
(1.青岛大学自动化学院 青岛 266071;2.青岛大学未来研究院 青岛 266071)

摘 要: 基于人体骨骼数据分析人体行为的方法可解释性强,在基于视觉的人体行为分析研究中具有明显优势。但视角干扰及目标遮挡严重影响人体骨骼关节点的标定。本文提出了一种在人体结构约束条件下的基于人体姿态特征的人体骨骼关节点估计算法,并根据骨骼数据识别人体行为。首先根据人体运动的稳态趋势和暂态变化,基于决策树和加权线性回归分别建立特征提取模型,对缺失或混淆的关节点进行估计。然后设计了一个结合轻量级时间卷积和注意力图卷积的行为识别网络模型,针对行为样本的时间尺度优化模型。在 NTU RGB+D 60 数据集中建立遮挡情况进行实验,准确率分别达到 90.28%(CV)与 81.95%(CS),且在 UTD-MHAD 数据集中达到 98.2%,均优于现有方法。

关键词: 行为识别;人体姿态估计;时间卷积;图卷积;运动估计
中图分类号: TP391.9 **文献标识码:** A **国家标准学科分类代码:** 520.52020

Human action recognition based on joint motion estimation

Li Zhihan[1,2] Liu Yinhua[2] Xie Ruikang[2] Shan Liang[2]
(1. Automaion Institute, Qingdao University, Qingdao 266071, China;
2. Institute for Future, Qingdao University, Qingdao 266071, China)

Abstract: The method of analyzing human behavior based on human skeleton data is highly interpretable and has obvious advantages in the research of human behavior analysis based on vision. However, viewing angle interference and target occlusion seriously affect the calibration of human skeleton joints. This paper proposes a human skeleton joint point estimation algorithm based on human pose features under the constraints of human structure, and recognizes human behavior based on skeleton data. Firstly, according to the steady-state trend and transient changes of human motion, feature extraction models are established based on decision tree and weighted linear regression, respectively, to estimate missing or confused joint points. Then, an action recognition network model combining lightweight temporal convolution and attention graph convolution is designed to optimize the model for the time scale of action samples. The occlusion condition was established in the NTU RGB+D 60 dataset for experiments, and the accuracy rates were 90.28%(CV) and 81.95%(CS), respectively, and 98.2% in the UTD-MHAD dataset, which were better than those of the existing methods.

Keywords: action recognition;human pose estimation;temporal convolution;graph convolution;motion estimation

0 引 言

随着视频监控的普及与广泛使用,人体行为识别(human activity recognition, HAR)在公共安全监控,人机交互,医疗看护,交通监管等领域都有良好表现[1]。检测场景中通常存在遮挡、模糊、多物体等干扰因素,设计一种准确率高,实时性强的人体行为识别方法是该领域亟需解决的问题。

基于 RGB 图像的行为识别方法[2-3],通过提取人体轮廓、质心等特征作为输入,辅以光流[4]、时序[5]等信息进行识别。但受到环境中障碍物和人体运动产生的光线变化,所提取的特征信息中包含扰动因素,对行为分类带来困难。基于人体骨骼关节点的行为识别方法[6-7]解决了部分问题,首先通过姿态估计算法[8-9]从图像中提取人体关节点,传统基于骨骼点的方法直接将关节点的坐标数组进行编码作为特征信息,如 2stream-3DCNN 模型[10]分别从空间和时间

收稿日期:2022-06-20
* 基金项目:国家重点研发计划重点专项(2020YFB1313600)资助

两个流对关节坐标编码；时间卷积神经网络（temporal convolutional networks，TCN）[11]使用二维卷积核对数据进行流式处理。

这类方法将关注点集中在人体姿态上，减少了复杂背景环境的干扰。但是没有考虑到关节点在人体骨骼中的自然联系，以及关节点坐标变化的相互依赖关系。而使用图卷积网络（graph convolutional networks，GCN）[12]可以有效表述与提取这类特征。Yan 等[13]提出时空图卷积网络（spatial temporal graph convolutional networks，ST-GCN）模型，对单帧关键点坐标数组建立邻接矩阵，通过图卷积提取空间特征，设置滑动窗口融合连续帧的时序特征，但仅考虑了可见的关节点联系，忽略了运动特征。对此，Shi 等[14]提出自适应图卷积（two-stream adaptive graph convolutional networks，2s-AGCN），设置了一个动作矩阵以学习所有关节之间的潜在关系，邻接矩阵与其相加得到更深层的关节点联系。而 Liu 等[15]提出一种跨时空信息流模型，设计了一个多尺度权重计算方法，解决了时空图卷积模型中偏权重问题。胡锦林等[16]在空间图卷积和时间卷积之间加入全局注意力模块，增加对全局特征信息的学习能力，并构造了一个六类行为的遮挡数据集进行验证。

综上所示，基于人体骨骼的行为识别方法具有较好的可解释性，但是现有的方法主要基于数据集中给出的标准骨骼点数据进行实验，没有充分考虑到实际情况中，由于视角干扰及目标遮挡产生的数据缺失问题。本文基于决策树建立稳态特征回归模型，结合线性回归模型提取的暂态特征对人体骨骼关节点进行估计，在数据方面克服遮挡和轮廓模糊问题。然后以关节点序列数据为输入，设计了一个行为识别网络模型。改进的轻量级时间卷积模块（light-TCN）能有效扩大模型的感受野，学习动作样本的自相关性，同时嵌入注意力机制的图卷积模块，在不同的动作样本中为关节点动态分配权重，以增强关节和肢体的层次特征。最后通过卷积完成样本的多分类，并在公开数据集上进行验证。

1　姿态估计模型

本文提出的方法整体框架如图 1 所示，输入一组视频流数据，通过姿态估计和运动估计模块得到较准确平滑的人体关节点数据，输入到所设计的行为识别模型中进行分类识别。

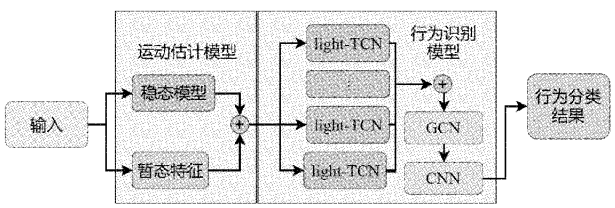


图 1　模型框架

1.1　人体姿态模型

人体姿态的关节位置决定了人类运动的几何结构，在行为识别任务中，主要关注的目标是人的姿势，其空间位置相对并不重要。为了更有效地识别人体行为，需要从复杂的背景中获取人体关节点信息。以 coco17 关键点模型为例，关键点序号与对应的身体关节部位如表 1 所示，手掌和脚掌的旋转不影响检测结果，这些部位被简化为单个关键点。

表 1　关节点名称与对应序号

序号	关节名	序号	关节名	序号	关节名
0	鼻子	6	左肘	12	左膝
1	脖子	7	左手腕	13	左脚踝
3	右肘	9	右膝	15	左眼
4	右手腕	10	右脚踝	16	右耳
5	左肩	11	左胯	17	左耳

单帧图像中，一组关键点被标记为 P0～P17，每个点的空间位置表示为 $\boldsymbol{x}_i=[x_{ix},x_{iy},x_{iz}]$，式中 $i\in[0,17]$，人体的运动可以用一个向量描述如(1)所示。

$$\boldsymbol{X}_i=[x_0,x_1,x_2,\cdots,x_{17}]^T\in R^{18\times 3} \tag{1}$$

式中：$\boldsymbol{X}_i$ 连续。以摔倒动作为例，一个完整的跌倒过程应包括以下状态：行走/站立、跌倒、躺在地上、爬起。从单帧图像中很难判断是否发生摔倒事件，而从视频流中可以观察到动态的过程。将一个动作定义成长度为 T 的连续帧如下所示。

$$\boldsymbol{x}_i(t)=[x_{ix}(t),x_{iy}(t),x_{iz}(t)] \tag{2}$$

式中：$1\leqslant t\leqslant T$，$t\in Z$，并且 $\boldsymbol{x}_i(t)$ 是连续的。计算位置向量对时间的导数，可以得到每个关键点的速度和方向如下：

$$\dot{\boldsymbol{x}}_i(t)=[\dot{\boldsymbol{x}}_{ix}(t),\dot{\boldsymbol{x}}_{iy}(t),\dot{\boldsymbol{x}}_{iz}(t)] \tag{3}$$

式中：$\dot{\boldsymbol{x}}_i(t)$ 是连续的。在姿势变化的过程中，肢体随着关节发生运动，可以用相邻关节的向量表示。关节点 $\boldsymbol{x}_j$ 和点 $\boldsymbol{x}_i$ 之间的肢体 $\boldsymbol{v}_{ij}$ 如下：

$$\boldsymbol{v}_{ij}(t)=\boldsymbol{x}_j(t)-\boldsymbol{x}_i(t) \tag{4}$$

式中：i 和 j 分别代表两个相邻的关节。当人体的姿势发生变化时，肢体的变化如下：

$$\dot{\boldsymbol{v}}_{ij}(t)=\dot{\boldsymbol{x}}_j(t)-\dot{\boldsymbol{x}}_i(t) \tag{5}$$

式中：$\dot{\boldsymbol{v}}_{ij}(t)$ 是连续的。各关节点和肢体的位置和变化趋势可以根据上述公式进行计算，在不同的姿态下具有不同的位置特征，这些特征与标准动作特征 $\boldsymbol{X}_{im}$ 进行比较，计算欧式距离如下：

$$dis_m=\sqrt{\sum_{i=1}^{i}(\boldsymbol{X}_{im}(i)-\boldsymbol{X}_i(i))^2} \tag{6}$$

不同的动作有不同的帧长度，本文按照等时间间隔 F 进行采样，以保留动作的长度信息。较长的采样间隔会导致关键点的过度变化，太短的时间间隔会增加计算量，降低检测效率。在长度为 T 的视频流 $\boldsymbol{S}$ 中，总共有 $T/F=N$

帧图像,所提取的特征序列如下:

$$\boldsymbol{S}=[X_i(k_1),X_i(k_2),\cdots,X_i(k_n)]^{\mathrm{T}},i\in[0,17] \quad (7)$$

式中:k 离散,$k_1,k_2,\cdots,k_N$ 表示帧在视频流中的序列,T 的值是任意实数,所以 $\boldsymbol{S}$ 可能包含一个或多个动作。当身体重叠、遮挡、倒置或图像模糊时,都会导致关键点生成丢失和错误。为减少误差点引起的误差,对关键点进行过滤,如下所示。

$$\boldsymbol{X}_i(k_n)=\begin{cases}\boldsymbol{X}_i(k_n), d_{ij}\leqslant|\boldsymbol{X}_i(k_n),\boldsymbol{X}_j(k_n)|\leqslant D_{ij}\ and\ v_{ij}(k_n)\in[0,R_{ij}]\\ 0,\ None\ \ meet\ \ the\ \ requirements\end{cases} \quad (8)$$

式中:d_{ij} 和 D_{ij} 分别表示关节点 i 和点 j 之间的最小和最大距离,R_{ij} 为点 i 和点 j 之间肢体的自然长度。这一步主要是确定关键点是否存在及其大致位置,不需要精确计算每个关节坐标。

1.2　运动估计模型

当人体姿态发生变化时,运动中的关键点遵循两个条件:一是场景中人体所包含的像素不会在帧之间发生变化,二是像素位置不会随时间发生剧烈变化。关节点的运动存在长期趋势,如人体的位置变化,肢体的周期性运动等称为稳态特征;而运动中发生在短时间内的姿态变化,如挥手抬腿等称为暂态特征,结合两种特征,在约束条件下建立了一个关节点轨迹预测模型,其结构如图 2 所示。

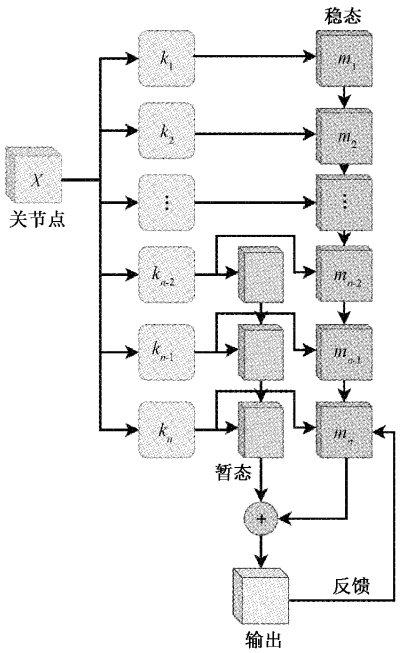


图 2　运动估计模型

基于决策树模型建立关节点数据的稳态特征预测模块,该模块由多个弱学习器组成,分配不同的权重组合成为一个强学习器。稳态特征模块的输入包括当前时刻与之前所有时刻数据,例如第 n 个输入样本 $\{\boldsymbol{X}(k_1),\cdots,\boldsymbol{X}(k_n)\}$ 是一个 $n\times36$ 维数组,计算方法如下:

$$f_g^n(x)=\sum_{i=1}^{n}\omega_g^i m_g^i(x)=f_g^{n-1}(x)+\omega_g^n m_g^n(x) \quad (9)$$

式中:每一个样本生成对应的决策树模型,如 $m_n(x)$ 表示第 n 个模型,为每个模型赋予权重 ω_g^n,所有基础模型的预测值累加,得到最终的预测值。

稳态特征模型依靠于长时间尺度信息进行训练,对运动趋势的估计较为准确,但会随着数据增加出现欠拟合现象。以本文使用的数据集为例,每秒采样 30 帧图像,人体运动中动作的平均变化时长在 3~5 帧,长时间尺度变化趋势已在稳态模型中提取,所以选择最新的连续三组数据表示暂态特征,建立加权线性回归模型:

$$f_l^n(\theta)=\sum_{i=n-2}^{n}\omega_l^i(f_l^i(x)-f_t^i(x))^2 \quad (10)$$

式中:预测值 $f_l^i(x)=\theta^{\mathrm{T}}x_i$,权重 ω_l^i 通过计算预测值与真实值的距离得到,权重函数如下所示。

$$\omega_l^i=\exp\left(-\frac{(f_l^i(x)-f_t^i(x))^2}{2k^2}\right) \quad (11)$$

式中:参数 k 控制权重的变化率,设置为 0.01。暂态特征生成短期预测值,与稳态特征预测值共同决策,为训练运动估计模型建立目标函数。

$$L=\sum_k l(f_g^n(x),f_l^n(x))+\sum_k l(f_l^n(x),f_t^n(x)) \quad (12)$$

将两者相对真实值的误差计算损失,并分别反馈给稳态模块和暂态模块。

$$\omega_l/\omega_g=\sum_k l(f_g^n(x),f_l^n(x))/\sum_k l(f_l^n(x),f_t^n(x)) \quad (13)$$

式中:$\omega_l+\omega_g=1$,设置一个输出权重用于结合两个特征模块的关节点估计结果,根据预测值与真实值的误差进行权重比例调整。

2　行为识别模型

本文提出的动作识别模型主要分为两个部分:时序卷积模块和空间卷积模块,整体结构如图 3 所示。时序卷积模块提取连续骨骼关节序列数据的时序特征,以矩阵的形式输出到空间特征提取模块,再使用图卷积对特征信息信息进行处理,最后通过卷积层得到动作分类结果。

2.1　时序卷积模块

为了有效提取视频流数据中存在隐含的时序特征,本文基于 TCN 构建了时序卷积模块。单个 light-TCN 模块结构如图 4(b)所示,在 TCN 模块的基础上,减少了一个卷积层,并引入了注意力机制。以时间序列输入的数据分别通过 tanh 激活函数和 sigmoid 激活函数(σ),同时保留了残差机制(一个 1×1 的卷积层),从输入中提取原始特征融合输出,以缓解梯度消失和爆炸问题。最后合并每个时刻 light-TCN 模块的输出特征,高层会根据低层提取的特征

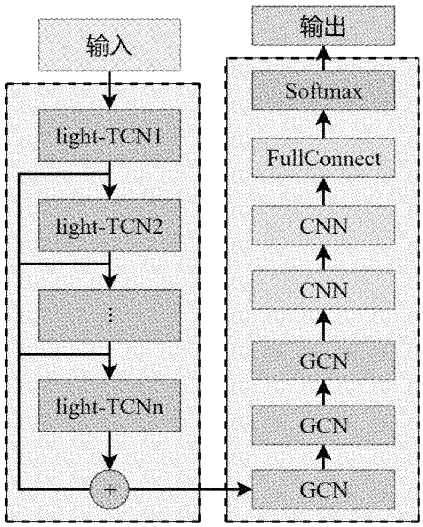


图 3　行为识别模型框架

信息进一步提取时序信息，最后将合并结果传输到图卷积模块中。

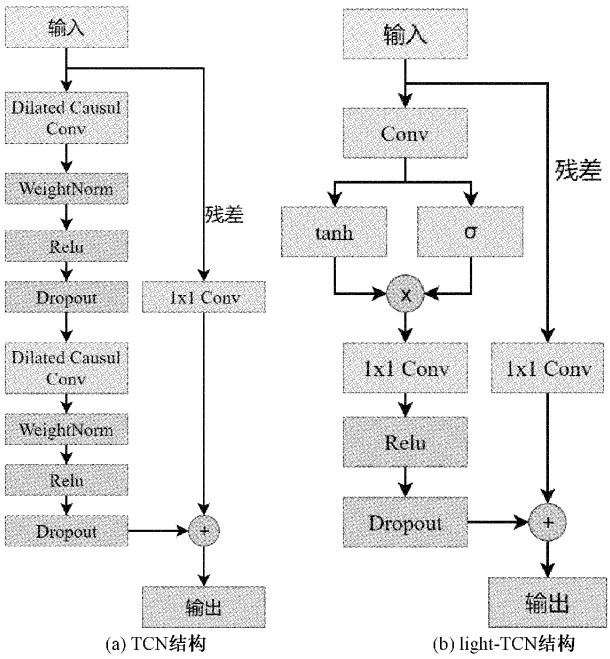


图 4　时序卷积模块

以三层时序提取网络模块为例，网络内部信息传输如图 5 所示，这种结构有两个特性：因果卷积（causal convolution）和膨胀卷积（dilated convolutions）。因果卷积是单向结构的时间约束模型，即对于当前时刻 t 的输出值，只依赖之前的输入与其通过卷积得到的中间特征状态。高层的感受野与网络层数呈线性关系，对于超长序列，网络必须很深，才能捕捉到足够长的历史信息。膨胀卷积是一种间隔采样机制，采样率受图中的 d 控制。隐含层的第一层对每个时刻的输入采样，第二层每两个点采样一个作为输入，第三层每四个点，以此类推。这种机制能使用较少的参数和层数实现更大的感受野。

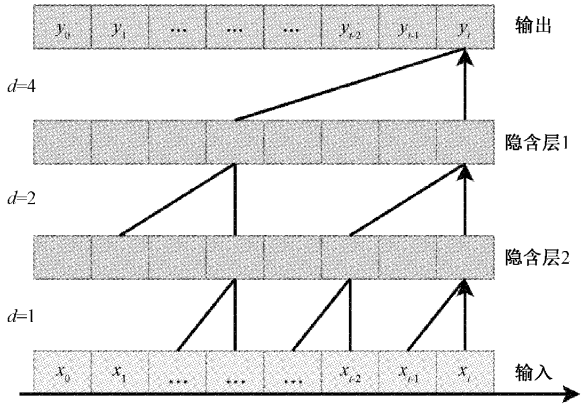


图 5　时序输入示例

2.2　空间卷积模块

在卷积神经网络中，通过共享的卷积核遍历图像，每个区域中像素点的值与卷积核对应的值相乘，然后求和作为该区域的特征值。在这个过程中，包含背景环境的像素同样加入卷积运算中，提取的特征信息存在干扰信息。从人体姿态模型中可以看出，动作识别只与图像中人体的姿态有关，而与背景环境无关。

姿态估计得到的坐标中主要包含两种信息：关节点的坐标，相邻关节点计算得到肢体信息。根据图卷积[12]理论，在视频流的某一帧，人体姿态可以用集合 $\boldsymbol{G}=(\boldsymbol{E},\boldsymbol{V})$ 表示，N 为关节点个数，其中 $\boldsymbol{E}\in R^{N\times N}$ 表示肢体的集合，$\boldsymbol{V}\in R^{N\times 3}$ 表示关节点的集合，每个点包含 x,y,z 三维坐标值。同时为了有效表示点之间的链接关系，设定了只包含 0 和 1 的邻接矩阵 $\boldsymbol{A}$（adjacency matrix）。$\boldsymbol{A}$ 包含 $N\times N$ 个元素，这些元素的值可以表示为：

$$\boldsymbol{A}(i,j)=\begin{cases}1, & i\Leftrightarrow j\\ 0, & \text{其他}\end{cases} \tag{14}$$

式中：$\Leftrightarrow$ 表示关节点 i 与 j 在人体模型中相邻。单帧图像中，图卷积输出可以定义为：

$$\boldsymbol{f}_{out}(x)=\sum_{k-1}^{K}\sum_{w-1}^{K}\boldsymbol{f}_{in}(p(x,h,\omega))\cdot\omega(h,\omega) \tag{15}$$

式中：$\boldsymbol{f}_{in}$ 为输入，卷积核大小为 $K*K$，采样函数 p 对 x 周围 K 邻域的像素点采样，然后权重函数 ω 与采样区域卷积，得到邻域像素点与采样中心点的关系以人体姿态为输入时，采用中心点是 V，邻域为：

$$B(V_i)=\{V_j\mid d(V_j,V_i\geqslant D)\} \tag{16}$$

式中：d 表示关节点距离采样中心点的最小距离，D 即邻域的大小。图卷积在空间的顺序取决于根节点向周围的标记顺序，将每个关节点 V_i 的邻域 B 划分成固定数量 N 的子集 l，其中每个子集都具有相同的标签，即 $\sum_{1}^{N}l(V_i)=B(V_i)$，可以得到新的权重函数。

$$\omega(V_i,V_j)=\omega'(l_i(V_j)) \tag{17}$$

将关节点的权重和采样函数代入图卷积公式，可以得

到基于关节点的图卷积表达式。

$$\boldsymbol{f}_{out}(V_i)=\sum_{V_j\in B(V_i)}\frac{1}{Z_i(V_j)}*\boldsymbol{f}_{in}(p(V_i,V_j))*\omega(V_iV_j) \tag{18}$$

式中:$Z_i(V_j)=|\{V_k\mid l_i(V_k)=l_i(V_j)\}|$ 等于相应子集的基数。得到单帧图卷积的计算方法之后,由于视频流中关节点在连续帧之间运动,对包含时间序列的骨架序列图卷积进行定义,将邻域的概念也包含时间上前后帧的相邻关节点。

$$B(V_{ti})=\{V_{ti}\mid d(V_{ti},V_{tj})\leqslant K,\mid q-t\mid\leqslant\mid\Gamma/2\mid\} \tag{19}$$

式中:Γ 表示时间卷积核的大小,时间图卷积同样需要将邻域划分子集,考虑到时间轴有序,对上文中子集公式进行修改。

$$l_{st}(V_{qj})=l_{tj}(V_{tj})+(q-t+[\Gamma/2])*K \tag{20}$$

图卷积模块使用多个 GCN 层提取特征,单个 GCN 模块结构如图 6 所示,该模块的输出计算公式为:

$$\boldsymbol{f}_{out}=\sum_{k}^{K_v}\boldsymbol{W}_k\boldsymbol{f}_{in}(\boldsymbol{A}_k+\boldsymbol{C}_k) \tag{21}$$

式中:$\boldsymbol{A}$ 为关节邻接矩阵,$\boldsymbol{W}_k$ 是图卷积权重,$\boldsymbol{f}_{in}$ 即输入共分为 4 个路径,其中 θ 和 φ 为高斯嵌入函数的两个分支,分别使用 1×1 卷积对输入进行线性变换,得到每个样本的特征图 $\boldsymbol{C}_k$,计算公式为:

$$\boldsymbol{C}_k=soft\max(f_{in}^{T}w_{\theta k}^{T}w_{\varphi k}\boldsymbol{f}_{in}) \tag{22}$$

式中:$w_{\varphi k}$ 和 $w_{\theta k}$ 表示 θ 和 φ 的权重,使 GCN 模块在识别不同动作时能够关注不同的关节,从而可以为每个关节动态分配不同的注意力权重,两个分支的特征融合之后,由 softmax 函数归一化至[0,1]。GCN 模块中同样使用了残差机制,以防止梯度爆炸。

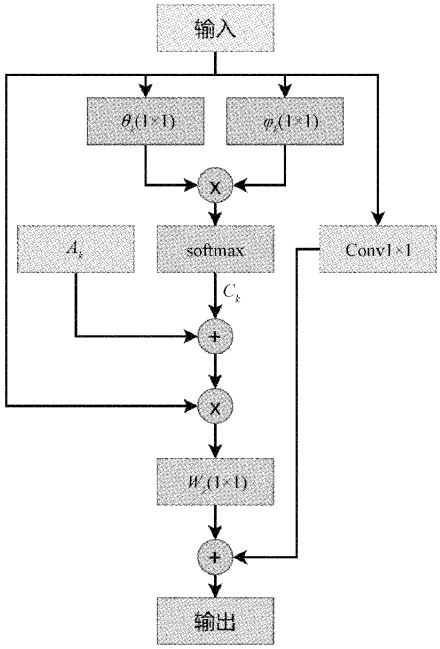


图 6 GCN 结构

GCN 模块提取的特征通过卷积层和 softmax 层进行分类,整个网络模型输出一个一维数组,数组长度与动作类别数量相等,数组中每个元素表示该动作的预测概率。

3 实 验

3.1 数据集

为了验证所提出方法的有效性,本文收集了两个网络公开数据集:

1) UTD-MHAD[17] 包含 8 位不同受试者执行的 27 个不同动作的 RGB 视频数据,每个动作重复 4 次,共 861 个 RGB 视频序列。

2) NTU-RGB+D 60[18] 包含 60 个种类的动作,共 56 880 个视频样本。按照不同的方式划分该数据集,可以分为按照人物 ID 来划分训练集和测试集的 Cross-Subject (CS),和按相机来划分的 Cross-View(CV)。

3.2 实验设置

考虑到输入数据来自于不同角度和场景的摄像头,进行归一化,变换坐标使动作主体位于视野中央,得到尺寸为 $C\times T\times V$ 的多维 joints 数据,其中 C 为坐标维数、T 为帧数、V 为点数 18。将关键点的连接关系输入邻接矩阵中,即得到骨骼数据。实验基于 windows 系统、Nvidia2060 显卡进行,设置训练轮数为 100,batchsize 为 64,模型学习使用 Adam 优化器,模型学习率为 0.001,评价指标为准确率(accuracy)和损失(loss)。

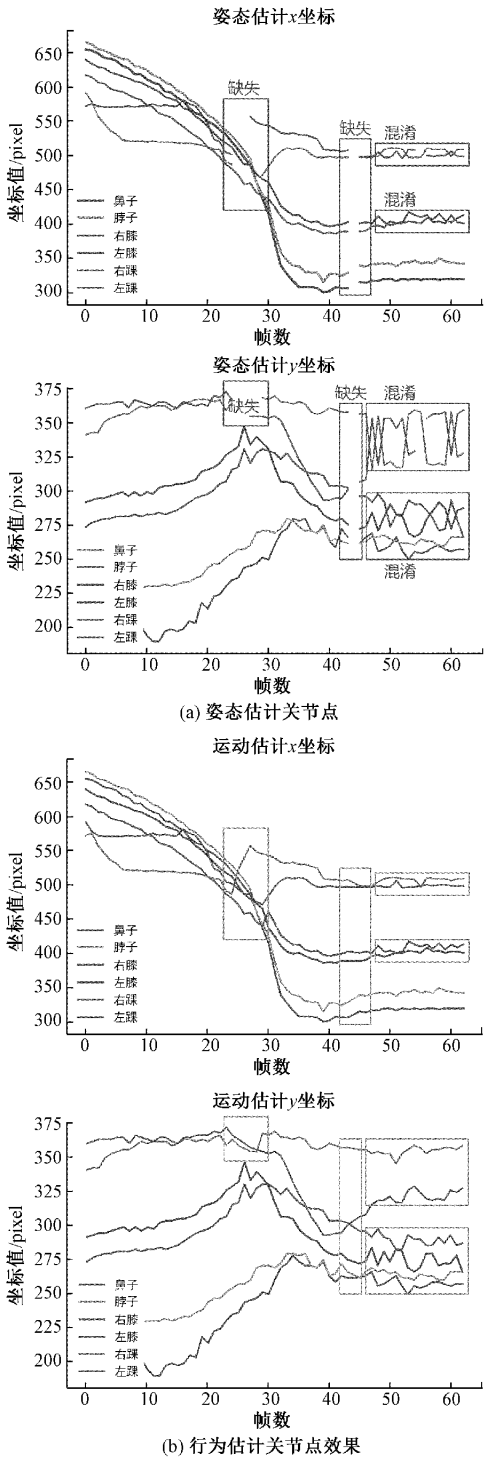
3.3 轨迹预测效果

取一段动作样本为例,直接使用姿态估计得到的结果如图 7(a)所示。主要存在 3 种错误现象:自身遮挡、环境遮挡和关节点混淆,在运动过程中,受角度、姿势等影响,部分关节点会与人体其他部位暂时检测重叠,如第 25 帧出现的部分曲线中断。在第 43 帧则出现整个人体被遮挡的现象,而在第 45~49 帧和第 60~64 帧时,身体左右的对应部位几乎完全混淆。由图 7(b)可看出,经过预测模型得到的轨迹可清晰区分,且不含缺失值。这种方法也可以用于矫正正确的关节点,选择更平滑的运动轨迹作为特征信息作为输入神经网络模型。

3.4 行为识别模型消融实验

1) 不同感受野的效果对比

在时序特征提取模块中,不同数量的 light-TCN 模块堆叠会产生不同的感受野,为了探究感受野对 模型性能的影响,在 NTU RGB+D 60 数据集 CV 中随机选择了部分数据进行对比实验。不使用 light-TCN 时等同于输入单帧图像,不提取时序特征。随着层数增大,单个样本包含的图像帧指数增加,由于数据集中最短的动作样本是 32 帧,平均长度 82.9 帧,设置了五层到八层 light-TCN 进行对比,感受野分别为 32、64、128、256。感受野大于样本帧数时,将样本重复增添至感受野大小,实验结果如表 2 所示,随着层数增加,准确率逐步上升,感受野与样本大小接近时得到



(a) 姿态估计关节点

(b) 行为估计关节点效果

图 7　关节点预测效果

最优效果，而在提取长持续时间动作特征时，可以选择较大的层数。本文的实验中，所用的时序提取模块均为六层。

2) 不同特征综合模块对比

为了证明 GCN 模块对特征的综合识别能力，分别使用 CNN、LSTM、GCN+LSTM、GCN+CNN 作为特征综合模块，进行对比实验，实验结果如表 3 所示。使用 CNN 对特征进行分类时，由于 CNN 是对全局共享卷积层，没有考虑到不同关节的权重，导致对行为特征的判别性不强。考虑到 LSTM 适用于对时序特征提取，实验中仅使用 LSTM 进行特征提取和分类，由于骨骼数据序列中不同样本存在相同的动作，模型通过当前数据预测存在滞后性，效果不佳。TCN 捕获全局信息，GCN 融合关节和肢体特征，降低了特征的同化，得到较好的准确率。综合考虑选择了 light-TCN+GCN+CNN 作为本文的行为识别模型。

表 2　不同感受野的准确率

Blocks	Acc/%	Loss/%
None	77.32	34.20
4-light-TCN	82.60	20.58
5-light-TCN	87.10	14.96
6-light-TCN	**89.55**	10.88
7-light-TCN	88.36	12.40

表 3　不同卷积模块的准确率

模块	Acc/%	Loss/%
TCN+CNN	80.20	23.40
仅 LSTM	75.31	30.59
TCN+GCN+LSTM	90.85	11.58
TCN+GCN+CNN	**92.30**	**10.02**

3) 不同样本对比

为了验证对不同动作的识别程度，从 UTD-MHAD 数据集中选择了 13 个动作类别，分别为挥手 1、挥手 2、行走、拍手、双手交叉、投掷、推、拳击、慢跑、站起、坐下、站立，将模型对数据集各个类别的识别结果制作混淆矩阵。如图 8 所示，横轴为使用网络模型预测的人体行为类别，纵轴为数据的真实标签。从混淆矩阵中可以看出，模型能有效准确识别大部分类别，但是对于具有相似关节位置的动作，例如挥手和推、投掷等，以及包含相同姿态的动作，例如行走、坐下都包含站立姿态，仍然存在一定的混淆概率。

3.5　与其他方法比较

1) UTD-MHAD 数据集

为了评估模型在 UTD-MHAD 数据集上的识别准确率，与 4 个不同方法达到主流网络模型进行对比，结果如表 4 所示。在识别准确率方面，虽然 Khaire 的方法平均准确率最低[20]，但是在部分类别上其准确率能达到 100%，这说明多视觉线索的结合在增强人体特征的同时，也会引入更多的干扰信息。Zhao[21] 和 Cao[22] 仅使用关节点为输入，模型准确率高于使用 RGB 图像的方法，因为加入了深度信息作为关节点的第三维度，或者通过残差机制增强关节点内部联系。使用本文所提出的动作估计和行为识别模型，减少了数据中的信息损失，得到较好的准确率，优于上述方法。

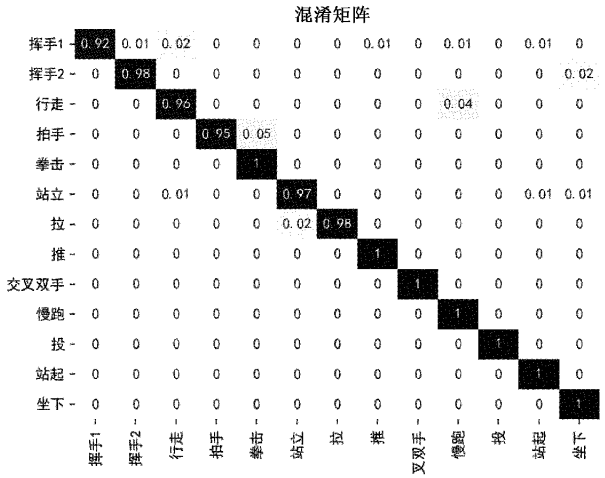


图 8 UTD-MHAD 数据集的混淆矩阵

表 4 UTD-MHAD 数据集与其他方法的对比

文献	模型	特征	准确率
Singh[19]	DMM+VGG	RGB 图像和深度图	0.957 4
Khaire[20]	结合多视觉线索的 ConvNets	RGB、深度图和关节点	0.946
Zhao[21]	贝叶斯 LSTM	2D 关节点	0.921
Cao[22]	线性跳门连接	2D 关节点	0.979
本文	Light-TCN+GCN	姿态估计关节点	0.935
本文	Light-TCN+GCN	运动估计关节点	**0.982**

2) NTU RGB+D 数据集

NTU RGB+D 数据集提供了标准的关节点数据,为了验证本文方法的有效性,对数据进行遮挡处理,随机清除部分关节点数据,即将其坐标数据置零,创建一个遮挡情况下的行为识别数据集,与现有的主流开源网络模型进行对比,其中包括基于 CNN 和 GCN 的深度学习算法,结果如表 5 所示。本文的方法在 CV 标准下准确率达到 90.28%,在 CS 标准下准确率达到 81.95%,均优于现有方法。基于 CNN 的方法准确率较低,这是由于 CNN 对特征信息的卷积是全局共享的,出现缺失信息时会对识别结果产生干扰,而 GCN 方法中数据有优先级,边缘信息的缺失对结果影响较小。

表 5 NTU RGB+D 数据集与其他方法的对比

Methods	CS/%	CV/%
2-stream-3DCNN[10]	61.52	68.28
ST-GCN[13]	69.84	76.58
2s-AGCN[14]	73.58	82.65
EfficientGCN[23]	80.29	88.52
本文	**81.95**	**90.28**

4 结 论

本文提出了一种在人体结构约束条件下的基于人体姿态特征的人体骨骼关节点估计算法,有效增加了信息量,提高了姿态特征的准确性。并构建了一个 light-TCN-GCN 神经网络模型用于人体行为识别。根据不同的数据长度改变感受野,灵活调整模型结构,图卷积根据人体骨骼关节固有的连接关系提取特征,提高对动作识别的准确率。在 UTD-MHAD 数据集和 NTU RGB+D 遮挡数据集上进行实验,结果表明,本文提出的模型具有较好的准确率,优于现有的行为识别方法,能有效克服数据缺失问题。由于数据使用的人体骨骼关节点呈离散状态,而忽略了动态特征,未来的工作主要融合动态特征如关节角度变化、运动速率等建立运动估计模型。

参考文献

[1] 邓淼磊,高振东,李磊,等.基于深度学习的人体行为识别综述[J].计算机工程与应用,2022,58(13):14-26.

[2] 吴伟,于嘉乐.分层双线性池化图像行为识别方法[J].电子测量与仪器学报,2021,35(3):152-157,DOI:10.13382/j.jemi.B2003449.

[3] 罗旭飞,崔敏,张鹏.基于骨骼的双支融合模型的人体行为识别[J].电子测量技术,2022,45(11):140-146,DOI:10.19651/j.cnki.emt.2208880.

[4] LADJAILIA A, BOUCHRIKA I, MEROUANI H F, et al. Human activity recognition via optical flow: decomposing activities into basic actions[J]. Neural Computing and Applications, 2020, 32(21): 16387-16400.

[5] 郑萌萌,钱慧芳,周璇.基于监控视频的 Farneback 光流算法的人体异常行为检测[J].国外电子测量技术,2021,40(3):16-22,DOI:10.19652/j.cnki.femt.2002392.

[6] 游伟,王雪.人行为骨架特征识别边缘计算方法研究[J].仪器仪表学报,2020,41(10):156-164,DOI:10.19650/j.cnki.cjsi.J2006750.

[7] LI Y, JI B, SHI X, et al. Tea: Temporal excitation and aggregation for action recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 909-918.

[8] CAO Z, HIDALGO G, SIMON T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172-186.

[9] 王泽杰,沈超敏,赵春,等.融合人体姿态估计和目标检测的学生课堂行为识别[J].华东师范大学学报(自然科学版),2022(2):55-66.

[10] LIU H, TU J, LIU M. Two-stream 3D convolutional neural network for skeleton-based action recognition [EB/OL]. 2017. [2022-06-01]. http://export. arxiv. org/pdf/1705. 08106. pdf.

[11] KIM T S, REITER A. Interpretable 3d human action analysis with temporal convolutional networks [C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017: 1623-1631.

[12] SPINELLI I, SCARDAPANE S, UNCINI A. Adaptive propagation graph convolutional network[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4755-4760.

[13] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Thirty-second AAAI Conference on Artificial Intelligence, 2018: 7444-7452.

[14] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12026-12035.

[15] LIU Z, ZHANG H, CHEN Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 143-152.

[16] 胡锦林,齐永锋,王佳颖. 基于时空图卷积网络的学生在线课堂行为识别[J]. 光电子·激光, 2022, 33(2): 149-156, DOI: 10. 16136/j. joel. 2022. 02. 0384.

[17] CHEN C, JAFARI R, KEHTARNAVAZ N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor [C]. 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015: 168-172.

[18] SHAHROUDY A, LIU J, NG T T, et al. Nturgb+d: A large scale dataset for 3d human activity analysis[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010-1019.

[19] SINGH R, KHURANA R, KUSHWAHA A K S, et al. Combining CNN streams of dynamic image and depth data for action recognition [J]. Multimedia Systems, 2020, 26(3): 313-322.

[20] KHAIRE P, KUMAR P, IMRAN J. Combining CNN streams of RGB-D and skeletal data for human activity recognition [J]. Pattern Recognition Letters, 2018, 115: 107-116.

[21] ZHAO R, WANG K, SU H, et al. Bayesian graph convolution lstm for skeleton based action recognition[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6882-6892.

[22] CAO C, LAN C, ZHANG Y, et al, Skeleton-based action recognition with gated convolutional neural networks [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29 (11): 3247-3257.

[23] SONG Y F, ZHANG Z, SHAN C, et al. Constructing stronger and faster baselines for skeleton-based action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, DOI: 10. 1109/TPAMI. 2022. 3157033.

作者简介

李志晗,硕士研究生,主要研究方向为图像处理,行为识别。

E-mail: 2020020580@qdu. edu. cn

刘银华,博士,副教授,主要研究方向为计算机视觉。

E-mail: liuyinhua@qdu. edu. cn

谢锐康,本科,主要研究方向为计算机领域。

E-mail: 2019204051@qdu. edu. cn

单良,本科,主要研究方向为计算机领域。

E-mail: 2019204046@qdu. edu. cn