智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

时空域融合的骨架动作识别与交互研究

钟秋波， 郑彩明， 朴松昊

引用本文:
钟秋波, 郑彩明, 朴松昊. 时空域融合的骨架动作识别与交互研究[J]. 智能系统学报, 2020, 15(3): 601–608.
ZHONG Qiubo, ZHENG Caiming, PIAO Songhao. Research on skeleton-based action recognition with spatiotemporal fusion and humanrobot interaction[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(3): 601–608.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202006029

DOI: 10.11992/tis.202006029

时空域融合的骨架动作识别与交互研究

钟秋波[1,2], 郑彩明[1], 朴松昊[3]

（1. 宁波工程学院 机器人学院, 浙江 宁波 315211; 2. 哈尔滨工业大学 机器人系统与技术国家重点实验室, 黑龙江 哈尔滨 150001; 3. 哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001）

摘　要: 在人体骨架结构动作识别方法中, 很多研究工作在提取骨架结构上的空间信息和运动信息后进行融合, 没有对具有复杂时空关系的人体动作进行高效表达。本文提出了基于姿态运动时空域融合的图卷积网络模型 (PM-STFGCN)。对于在时域上存在大量的干扰信息, 定义了一种基于局部姿态运动的时域关注度模块 (LPM-TAM), 用于抑制时域上的干扰并学习运动姿态的表征。设计了基于姿态运动的时空域融合模块 (PM-STF), 融合时域运动和空域姿态特征并进行自适应特征增强。通过实验验证, 本文提出的方法是有效性的, 与其他方法相比, 在识别效果上具有很好的竞争力。设计的人体动作交互系统, 验证了在实时性和准确率上优于语音交互系统。

关键词: 动作识别; 时空关系; 姿态运动; 时空域融合; 图卷积神经网络; 时域关注度; 自适应特征增强; 人体动作交互

中图分类号: TP312　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)03−0601−08

中文引用格式: 钟秋波, 郑彩明, 朴松昊. 时空域融合的骨架动作识别与交互研究 [J]. 智能系统学报, 2020, 15(3): 601–608.

英文引用格式: ZHONG Qiubo, ZHENG Caiming, PIAO Songhao. Research on skeleton-based action recognition with spatiotemporal fusion and human–robot interaction[J]. CAAI transactions on intelligent systems, 2020, 15(3): 601–608.

Research on skeleton-based action recognition with spatiotemporal fusion and human–robot interaction

ZHONG Qiubo[1,2], ZHENG Caiming[1], PIAO Songhao[3]

(1. Robotics Institute, Ningbo University of Technology, Ningbo 315211, China; 2. State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China; 3. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Temporal dynamics of postures over time is crucial for sequence-based action recognition. Human actions can be represented by corresponding motions of an articulated skeleton. Skeleton-based action recognition algorithm is used for studying motions of a body. Skeleton-based action recognition uses many methods, and research shows that most of them extract spatial and motion information separately from a skeleton structure and then combine them for further processing. However, this process is not able to efficiently deliver human motion features with complex temporal and spatial relationships. We propose a novel posture motion-based, spatiotemporal fused graph convolution network for skeleton-based action recognition. First, we define a local posture motion-based time attention module , which is used to constrain the disturbance information in temporal domain and learn the representation of motion posture features. Then, we design a posture motion-based, spatiotemporal fusion module. This module fuses spatial motion and temporal attitude features and adaptively enhances the skeleton joint features. Extensive experiments have been performed and the results verified the effectiveness of our proposed method. The proposed method has competitive performance, and it is concluded that the human–robot interaction system based on action recognition is superior to the speech interaction system in real-time and with respect to accuracy.

Keywords: action recognition; temporal and spatial relationships; posture motion; spatiotemporal fusion; graph convolution network; temporal attention; adaptive feature enhancement; human–robot interaction

随着人工智能技术的发展, 以人为核心的视觉人机交互技术的关键在于理解人类活动[1]和社会行为[2]。因此, 动作识别在人机交互领域具有

收稿日期: 2020−06−17.

基金项目: 国家自然科学基金项目 (61203360, 61502256); 浙江省自然科学基金项目 (LQ12F03001).

通信作者: 钟秋波. E-mail: zhongqiubo@nbut.edu.cn.

重要作用[3]。人体动作识别主流的两种方法是基于 RGB 和基于人体骨架序列。基于 RGB 的方法充分利用图像数据，在识别效果上可以获得较高的性能。这种方法通常需要处理图像上的每一个像素点来提取特征，高成本的计算资源无法满足实时处理的条件，不同照明条件和背景噪声对该方法影响较大。由于人体骨架的关节数量有限，且只有几十个，所耗费的计算资源要低，实时性有保障，对动态环境和复杂背景的适应性强。微软的 Kinect、OpenPose[4] 以及 CPN[5] 等高精度人体姿态估计也适合用于提取人体骨架。

已有的动作识别工作中，龚冬颖等[6] 通过随机森林算法和信息熵分析关节点判别力，自适应选择动作表示特征，解决了多特征融合对个别特征的分类缺陷。姬晓飞等[7] 将动作交互过程分阶段分别提取感兴趣区域的 HOG 特征，并加权融合每个阶段的分类概率，解决交互区域难分割的问题。庄伟源等[8] 选择高判别力的关键肢体的角度直方图作为特征解决多尺度以及相似动作的分类。徐志通等[9] 提出了联合加权稀疏重构轨迹与幅值方向直方图熵的方法来检测异常行为。吴云鹏等[10] 提出了基于流密度的聚类算法和多重邻接关系模型分别识别局部和全局模式解决复杂视频下的多人行为问题。陈婷婷等[11] 采用慢特征分析提取不变量特征，这种方法提取的特征具有很强的区分力。丁重阳等[12] 提出了时空权重姿态运动特征的方法，对骨架不同关节点分配权重，并引入傅里叶时间金字塔算法和动态时间规整算法对时序运动建模分析。莫宏伟等[13] 提出改进的 Faster R-CNN 人体行为检测模型，将在线难例挖掘算法和批量规范化算法与 Faster R-CNN 算法相结合有效解决小样本难例的问题。谢旋等[14] 使用 CNN 提取深度特征输入到 LSTM+Inception V3 网络模型提取全局和局部的相关信息，完成双人交互识别与预测。谢昭等[15] 提出了双流时空关注度长短时记忆网络模型抑制时空背景信息。王传旭等[16] 使用双流 TSN 网络提取全局和局部特征，并分别连接 LSTM 网络捕获视频中的长期依赖关系，解决复杂环境下的多视觉信息和上下文信息融合问题。郑兴华等[17] 使用深度残差网络和递归神经网络对图像提取深度特征，通过 STRIPS 模型将深度特征表示为描述文档并使用规划器推导完整的动作序列解决行为数据缺失的情况。张冰冰等[18] 考虑视频局部特征的高阶统计信息，对一阶和二阶局部特征二阶聚合形成高阶全局特征表示，对空间和时间特征分布进行建模。

传统基于深度学习的方法是通过手工的方式将骨架序列构造为一组关节向量序列输入到 RNNs[19]，或者将表示骨架序列的语义图像输入到 CNNs[20] 中提取特征，最后对动作类别进行预测。然而关节向量序列和语义图像都不能很好地表达人体关节之间的相关性。最近，图卷积神经网络 (graph convolutional network,GCNs) 将卷积操作从二维的图像结构拓展到了图结构上，在很多应用上都有很好的表现。Yan 等[21] 首次使用 GCNs 应用在基于骨架结构的动作识别上，并提出时空图模型。Shi 等[22] 通过自适应学习的方法学习关节之间的连接关系构建非局部模块，将一阶的关节信息、二阶的骨骼信息以及运动信息等分别作为输入构建多流网络。Liu 等[23] 提出跨时空的多尺度聚合节点信息，有效消除了邻域节点的重要性。Peng 等[24] 使用骨架邻接图的高阶表征以及动态图建模机制寻找隐式的关节联系。Obinata 等[25] 通过跨时空的卷积方式对时空图进行时序建模，以提取运动关节的相关特征。这些方法都忽略了运动姿态和骨架关节特征在时空域上的特征融合。

现有的研究工作，不能有效地对时空图中的空间信息和运动信息进行融合实现端到端的训练。本文提出了基于姿态运动时空域融合的图卷积神经网络模型 (PM-STFGCN)，使用基于姿态运动的时空域融合模块 (PM-STF)，对运动姿态和骨架关节点在时空域上进行特征融合并自适应特征增强。针对时域内存在的大量干扰信息，定义一种基于局部姿态运动的时域关注度模块 (LPM-TAM) 进行有效的抑制。此外，与仅优化空间图卷积的方法相结合，本文提出的方法可以进一步提高识别的性能。本文在两个大型骨架数据集上进行实验，与常用方法进行比较。设计的人体动作交互系统与语音交互在实时性和识别准确率上进行对比。

1　相关工作

1.1　空间图卷积

时空图卷积神经网络模型[21] 构建的骨架时空图将节点之间的连接关系，用表示自身连接的单位矩阵 $\boldsymbol{I}$ 和表示相邻节点连接关系的邻接矩阵 $\boldsymbol{A}$ 表示。在单帧情况下，空间域上图卷积操作如下：

$$\boldsymbol{f}_{\text{out}}=\sum_{k=1}^{K}(\boldsymbol{\Lambda}_k^{-\frac{1}{2}}\boldsymbol{A}_k\boldsymbol{\Lambda}_k^{-\frac{1}{2}}\otimes\boldsymbol{M}_k)\boldsymbol{f}_{\text{in}}\boldsymbol{W}_k \tag{1}$$

式中：$\boldsymbol{f}_{\text{in}}$ 是输入维度为 (C_{in},T,N) 张量的特征图；C_{in} 是输入的特征通道数；T 是骨架帧的数量；N 表示骨架顶点的数量；$\boldsymbol{A}_k$ 是 $N\times N$ 的类邻接矩阵，$\boldsymbol{A}_k^{ij}=\{0,1\}$，其元素 $\boldsymbol{A}_k^{ij}$ 表示顶点 v_j 是否在顶点 v_i 的子集中，$\boldsymbol{A}_0$ 表示顶点自身，$\boldsymbol{A}_1$ 表示相比于根节点，距离骨架重心更近的节点，$\boldsymbol{A}_2$ 表示距离骨架重心更远的节点；$\boldsymbol{\Lambda}_k^{ii}=\sum_j(\boldsymbol{A}_k^{ij})+\alpha$ 是归一化对角矩阵，α=0.001；K 表示划分的子集个数，根据基于空间距离的划分规则[21]共有 3 个不同子集，即 K=3；$\boldsymbol{W}_k\in\mathbf{R}^{C_{\text{out}}\times C_{\text{in}}\times 1\times 1}$ 是 1×1 的卷积权重向量；C_{out} 是输出的特征通道数；$\boldsymbol{M}_k$ 是维度为 $N\times N$ 的空间注意力特征图，表示每个顶点的重要性；⊗是矩阵对应元素相乘，这意味着它只能影响与当前目标相连的顶点。

1.2　时间图卷积

文献 [22] 在时间域上提出了时间关注度模块，注意力系数计算如下：

$$\boldsymbol{M}_t=\sigma(g_t(\text{AvgPool}(\boldsymbol{f}_{\text{in}})))$$

式中：$\boldsymbol{f}_{\text{in}}\in\mathbf{R}^{C_{\text{out}}\times T\times N}$ 是输入的特征图，T 是指在时间上的长度；AvgPool 是平均池化；g_t 是 1×1 的卷积操作，权重系数为 $\boldsymbol{W}_g\in\mathbf{R}^{1\times C_{\text{in}}\times K_S}$，$K_S$ 是卷积核大小；σ 标记为 Sigmoid 激活函数；注意力图 $\boldsymbol{M}_{\text{t}}\in\mathbf{R}^{1\times T\times 1}$，表示在某个时刻的骨架图重要性程度。

文献 [21] 基于一种简单策略定义了时间维度上的图卷积，在式 (1) 中，在时间维度上使用核大小为 Γ×1 进行时间域上的图卷积。因此，在顶点 V_{ti} 上的采样区域为 $B(v_{ti})=\{v_{qi}\,|\,|q-t|\leqslant\lfloor\Gamma/2\rfloor\}$，其中 V_{ti} 表示时间序列 t 上的第 i 个顶点；Γ 是在时间维度上的核大小，在文献 [21] 中设为 9。

2　基于姿态运动的时空域融合图卷积

本文从人体运动具有复杂时空关系的角度出发，提出了基于姿态运动的时空域融合方法，融合时空域上的姿态运动和骨架关节特征并自适应增强特征。

2.1　姿态运动表示

姿态运动表达了在连续帧上对应关节顶点的运动信息。比如给定的第 t−1 帧的第 i 个关节顶点 $v_{t-1,i}=(x_{t-1,i},y_{t-1,i},z_{t-1,i})$ 和第 t 帧的第 i 个关节顶点 $v_{t,i}=(x_{t,i},y_{t,i},z_{t,i})$，其关节顶点运动姿态表示为 $m_{t,i}=(x_{t,i}-x_{t-1,i},y_{t,i}-y_{t-1,i},z_{t,i}-z_{t-1,i})$，$m_{t,i}$ 是第 t 帧第 i 个关节顶点的运动姿态。

2.2　基于局部姿态运动的时域关注度

对于在时域上存在大量的干扰信息，本文提出了基于局部姿态运动的时域关注度模块 (LPM-TAM)，如图 1 所示。骨架图上所有关节点运动信息表示为姿态运动，关节运动信息计算如下：

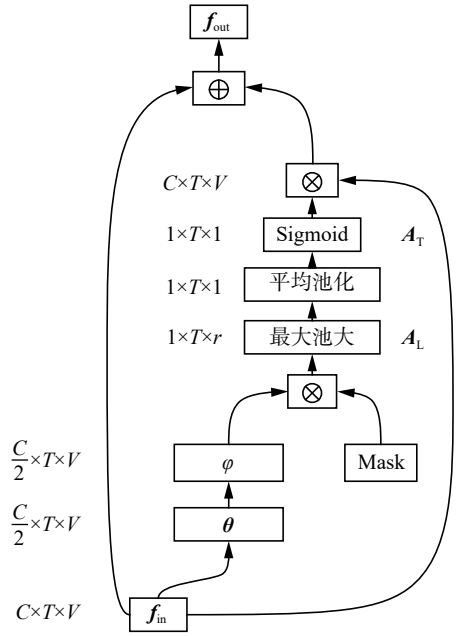


图 1　基于局部姿态运动的时域关注度模块

Fig. 1　Posture motion-based temporal attention module

$$\boldsymbol{M}=\sum_{t=1}^{T}\boldsymbol{\theta}_t(\varphi_t(\boldsymbol{f}_{\text{in}}^t))$$

式中：$\boldsymbol{f}_{\text{in}}^t\in\mathbf{R}^{C_{\text{in}}\times 1\times N}$ 是在 t 时刻的输入特征图，C_{in} 是输入的特征通道数，N 表示骨架顶点的数量；φ_t 用于提取特征图的运动信息；$\boldsymbol{\theta}_t$ 为降低特征图的通道数；$\boldsymbol{M}$ 是维度为 $\frac{C_{\text{in}}}{2}\times C_{\text{in}}\times T\times N$ 的运动特征图，特征通道为输入通道的一半，T 是骨架帧的数量。

人体的动作属于肢体运动，是部分或者全部的肢体进行运动。本文首先计算局部肢体在时域上的关注度 $\boldsymbol{A}_{\text{L}}$，通过局部肢体的关注度来表征动作序列在时域上的关注度 $\boldsymbol{A}_{\text{T}}$。局部肢体在时间维度上的注意力 $\boldsymbol{A}_{\text{L}}$ 的计算主要取决于局部感知域 P 内的运动特征。其中，$P=\{p_0,p_1,p_2,p_3,p_4\}$，即左手 p_0、右手 p_1、左腿 p_2、右腿 p_3 以及其他肢体部位 p_4。$\boldsymbol{A}_{\text{L}}\in\mathbf{R}^{1\times T\times r}$，$r$ 是指分解的肢体个数，本文中 r=5。基于局部姿势运动的时域关注度计算如下：

$$\boldsymbol{A}_{\text{T}}=\sigma(\text{AvgPool}(\text{MaxPool}(\textbf{Mask}\otimes\boldsymbol{M})))$$

式中：$\boldsymbol{A}_{\text{T}}\in\mathbf{R}^{1\times T\times 1}$ 是指在时间长度为 T 的时空图上，每帧动作骨架图的重要性程度；$\boldsymbol{M}$ 是顶点关节运动特征图；$\textbf{Mask}\in\mathbf{R}^{1\times 1\times N}$ 是局部肢体集 P 的掩码；⊗ 是矩阵对应元素相乘；σ 是指 Sigmoid 激

活函数。最后结果输出如下：

$$\boldsymbol{f}_{\text{out}} = \boldsymbol{f}_{\text{in}} \otimes \boldsymbol{A}_{\text{T}} \oplus \boldsymbol{f}_{\text{in}}$$

式中：输入特征映射以残差连接的方式乘以注意力特征图 $\boldsymbol{A}_{\text{T}}$ 实现自适应特征增强；⊕是指矩阵元素对应相加。

2.3　基于姿态运动的时空域融合

为了融合骨架关节和运动特征实现端到端训练，提出了基于姿态运动时空域融合模块 (PM-STF)，融合空域和时域运动特征并进行自适应特征增强。该模块在第 t 帧第 i 个顶点的时空域融合图卷积输出为

$$\boldsymbol{f}_{\text{out}}(v_i) = \boldsymbol{f}_{\text{in}}(v_i) + \sum_{v_j \in B'(v_i)} \frac{1}{Z_i(v_j)} m(v_j) w(l_j(v_j)) \qquad (2)$$

式中：$\boldsymbol{f}_{\text{in}}(v_i)$ 是顶点 v_i 的输入特征；$m(v_j)$ 是邻域顶点 v_j 的姿态运动特征；$l_j(v_j)$ 是指邻域顶点 v_j 划分子集映射的标签，使用基于空间距离划分规则划分为 3 个子集；w 为不同子集分配权重；$Z_i(v_j)$ 是划分子集的基数，用于归一化。

在实现过程中，式 (2) 转化为

$$\boldsymbol{f}_{\text{out}} = \boldsymbol{f}_{\text{in}} + \sum_{h=1}^{H} (\boldsymbol{\Lambda}_h^{-\frac{1}{2}} \boldsymbol{A}_h \boldsymbol{\Lambda}_h^{-\frac{1}{2}} \otimes \boldsymbol{A}_{\text{S}})(\boldsymbol{W}_m \boldsymbol{M}_{\text{in}}) \boldsymbol{W}_h \qquad (3)$$

式中：$\boldsymbol{M}_{\text{in}} \in \mathbf{R}^{\frac{C_{\text{in}}}{2} \times T \times N}$ 是姿态运动特征图；$\boldsymbol{W}_m \in \mathbf{R}^{C_{\text{in}} \times \frac{C_{\text{in}}}{2} \times 1 \times 1}$ 是 1×1 的卷积操作，增加运动特征图的特征通道；$\boldsymbol{W}_h \in \mathbf{R}^{C_{\text{out}} \times C_{\text{in}} \times 1 \times 1}$ 是 1×1 的卷积权重向量，C_{out} 是输出的特征通道数；$\boldsymbol{A}_{\text{S}} \in \mathbf{R}^{1 \times T \times N}$ 是空间注意力图，用于区分每个关节顶点的重要性；⊗是指矩阵对应元素相乘；$\boldsymbol{A}_h$ 是类邻接矩阵，$\boldsymbol{A}_h^{ij} = \{0,1\}$,其元素 $\boldsymbol{A}_h^{ij} = 1$ 表示顶点 v_j 在顶点 v_{i} 的子集 B_h' 中，否则 $\boldsymbol{A}_h^{ij} = 0$；$\boldsymbol{\Lambda}_h^{ii} = \sum_j (\boldsymbol{A}_h^{ij}) + \alpha$ 是归一化对角矩阵，α=0.001。

2.4　PM-STFGCN 的实现

本文的实现过程与仅优化空间图卷积的方法相结合。如图 2 所示，以 ST-GCN 为例，添加了本文提出的方法的实现过程。每一层时空图卷积层都包含了空间图卷积模块和时间图卷积模块。PM-STFGCN 模块在空间和时间图卷积模块之间附加本文的基于姿态运动时空域融合模块 (PM-STF)，融合时空图的空域位置特征和时域运动特征并在该模块输入端添加基于局部姿态运动的时域关注度模块 (LPM-TAM) 用于抑制时序上的干扰信息。S-GCN 和 T-GCN 分别是原本模型的空间和时间图卷积模块。最终，构建一个新的时空融合图卷积层 (PM-STGCN Block)。在实验过程中，本文将第一帧的运动信息设为 0。

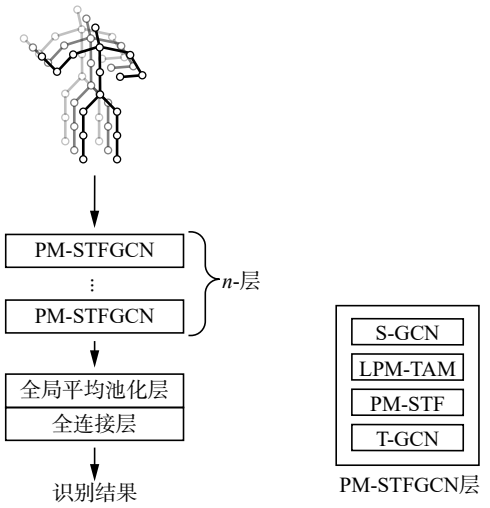


图 2　基于 ST-GCN 的姿态运动时空域融合图卷积模型

Fig. 2　Spatiotemporal graph convolution network-based spatiotemporal fusion graph convolution networks

2.5　人机交互的实现

本文设计了基于动作识别的人机交互系统，将动作交互实验应用在 NAO 机器人上，通过外置 Kinect v2 深度相机获取数据，交互实验算法流程如算法 1。

算法 1　人机动作交互

1) 获取 Kinect v2 的人体骨骼序列数据 X；

2) 将骨骼数据 X 输入到动作识别网络 (2s-AGCN+PM-STFGCN) 中；

3) 选取概率最大的预测动作类别 Y；

4) 通过字典将动作类别 Y 映射为数字指令 Y'；

5) 将数字指令 Y' 传输给 NAO 机器人；

6) NAO 机器人根据接收的数字指令作出指定的动作。

3　实验结果分析

3.1　数据集

NTU-RGB+D 是目前最大，应用最广泛的基于骨架的动作识别多模态数据集。它包含：60 个动作类，56 880 个视频片段，交叉对象 (CS) 和交叉视角 (CV) 两种训练基准。在交叉对象 (CS) 中，训练集包含 40 320 个动作样本，测试集包含 16 560 个动作样本。在交叉视角 (CV) 中，训练集包含 2 号和 3 号传感器拍摄的 37 920 个动作样本，测试集包含 1 号传感器拍摄的 18 960 个动作样本。本文在此基础上测试两个基准数据集的 top-1 精度。

3.2　消融研究

在 NTU-RGB+D 大型骨架数据集上进行了实验证明。将基于局部姿势运动的时域关注度模

块 (LPM-TAM) 和基于姿势运动的时空域融合模块 (PM-STF) 用 PM-STFGCN 表示。比较了 ST-GCN[21] 和 ST-GCN+PM- STFGCN、2s-AGCN[26] 和 2s-AGCN+PM-STFGCN。结果显示与原有模型相比, 后者性能得到了提升, 并分别验证 LPM-TAM 和 PM-STF 都是有效的。

如表 1 所示, 对于 ST-GCN[21] 和 ST-GCN+PM-STFGCN, PM-STFGCN 在 NTU-RGB+D 的 CS 和 CV 基准的 top-1 精度上分别提高了 3.9% 和 1.4%, 对于 2s-AGCN[26]和 2s-AGCN+PM-STFGCN, PM- STFGCN 在 CS 和 CV 基准的 top-1 精度上分别提高了 3.1%和 1.0%, 2s-AGCN+PM-STFGCN 在 CV 基准上表现最好。

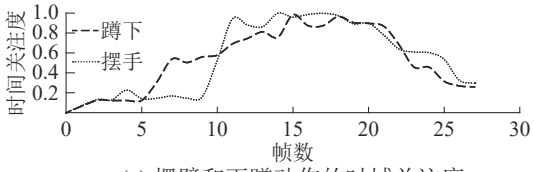
表 1　在 CS 和 CV 基准上的对比实验

Table 1　Ablation study on the CS and CV benchmark　%

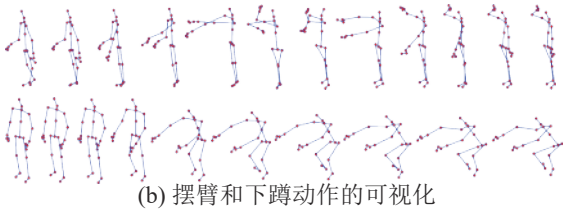
方法	精度	
	CS	CV
ST-GCN[21]	81.5	88.3
2s-AGCN[26]	88.6	95.2
ST-GCN+LPM-TAM	84.1	88.9
2s-AGCN+LPM-TAM	90.3	95.7
ST-GCN+PM-STF	84.7	89.2
2s-AGCN+PM-STF	91.1	95.9
ST-GCN+PM-STFGCN	85.4	89.7
2s-AGCN+ PM-STFGCN	91.7	96.2

为了验证所提出的 LPM-TAM 的有效性。在 ST-GCN[21] 和 2s-AGCN[26] 网络的时空图卷积层中只添加了 LPM-TAM, 并分别命名为 ST-GCN+LPM-TAM 和 2s-AGCN+LPM-TAM。结果如表 1 所示, ST-GCN+LPM-TAM 模型在 CS 和 CV 基准的 top-1 精度上分别提高了 2.6% 和 0.6%, 2s-AGCN+LPM-TAM 模型在 CS 和 CV 基准的 top-1 精度上分别提高了 1.7% 和 0.5%, 实验结果表明 LPM-TAM 是有效的。

对于时域关注度, 本文给出了人体骨架序列所对应的时域关注度示例。如图 3 所示, 图 (a)、(b) 分别是时序关注度、摆动手臂和蹲下为例的骨架序列的可视化。对于摆动手臂的时序, 模型更关注受试者摆动手臂的过程。对于蹲下的时序, 模型更关注受试者下蹲的过程。这可以有效地抑制时序上的不必要信息, 表明所设计模块的有效性和适应性。



(a) 摆臂和下蹲动作的时域关注度



(b) 摆臂和下蹲动作的可视化

图 3　摆动手臂和蹲下的时域关注度可视化

Fig. 3　Visualization of temporal attention map with arm swings and squat down as an example

本文验证了所提出的 PM-STF 有效性。在 ST-GCN[21] 和 2s-AGCN[26] 网络的时空图卷积层中只添加了 PM-STF, 并分别命名为 ST-GCN+PM-STF 和 2s-AGCN+PM-STF。结果如表 1 所示, ST-GCN+PM-STF 模型在 CS 和 CV 基准的 top-1 精度上分别提高了 3.2% 和 0.9%, 2s-AGCN +PM-STF 模型在 CS 和 CV 基准的 top-1 精度上分别提高了 2.5% 和 0.7%, 实验结果表明 PM-STF 是有效的。

3.3　与当前常用方法对比

将本文方法与当前常用方法的模型进行对比, 结果如表 2 所示。其中 2s-AGCN+PM-STFGCN 在 CV 和 CS 基准上取得非常好的表现。

表 2　本文方法与常用模型在 CS 和 CV 基准上的比较

Table 2　Comparison of our proposed method with current state-of-the-art methods on the CS and CV benchmark

方法	CS/%	CV/%	耗时/s
STA-LSTM[19]	73.4	81.2	—
CNN-based[20]	83.2	89.3	—
ST-GCN[21]	81.5	88.3	0.056
MS-AAGCN[22]	90.0	96.2	0.096
GCN-NAS[24]	89.4	95.7	—
2s-AGCN[26]	88.6	95.2	0.071
2s-AGCN+PM-STFGCN	91.7	96.2	0.082

在耗时上与 ST-GCN、2s-AGCN 和 MS-AAGCN 作比较。多流模态网络仅使用关节模态耗时代替整体算法的耗时且便于比较。本文提出的方法与原本模型相比在识别耗时上增加约 0.01 s,

增加的算法耗时对总体效果影响低。2s-AGCN+PM-STFGCN 平均识别速度达到 0.086 s/f, 平均每秒处理约 12 f 数据。相比于其他方法, 本文提出的方法优点在于识别效果的提升, 同时实验增加的算法耗时较少。

3.4　交互实验对比

为了进一步评估所提出的动作识别模型的鲁棒性, 将其应用于 NAO 机器人。表 3 为动作语义和交互动作之间的对应关系。

表 3　动作语义与交互动作对应关系

Table 3　Correspondence between action semantics and interactive action

动作语义	交互动作
挥手	坐下
向前走	前进
摸耳朵	站立
向后退	后退
抱手	左转弯
鼓掌	右转弯
屈膝	静止
双手抱头	关机

本文用准确率和实时性这两个指标进行实验验证, 将本文提出的动作交互与 NAO 机器人语音交互的响应时间进行对比, 记录了 50 次两种交互方式的准确次数, 验证了动作交互的可靠性。图 4、图 5 为以挥手和摸耳朵为例的动作交互场景。

图 4　以挥手为例的动作交互实验

Fig. 4　Action interaction with waving hand as an example

图 5　以摸耳朵为例的动作交互实验

Fig. 5　Action interaction with touching ear as an example

如图 6 所示, 动作交互相比于语音交互的准确次数更多。语音交互的平均识别精度为 85.50%, 动作交互的平均识别精度为 89.75%。与语音交互相比, 动作交互的识别精度提高了 4.25%。

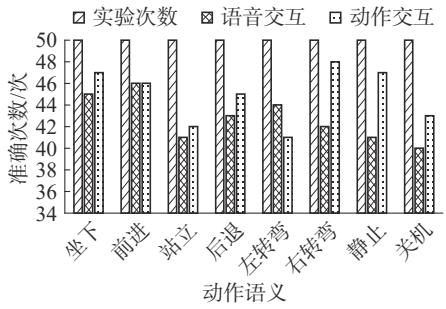


图 6　语音与动作交互的准确次数比较

Fig. 6　Comparison between accuracies of speech and action interaction

对语音和动作交互在响应时间上进行对比。如图 7 所示, 均为 10 次实验结果测量后的平均耗费时间。结果显示, 动作交互的响应时间大都短于语音交互。语音交互平均响应时间为 2.02 s, 动作交互平均响应时间是 1.71 s, 相比于语音交互, 其响应时间降低了 0.31 s。

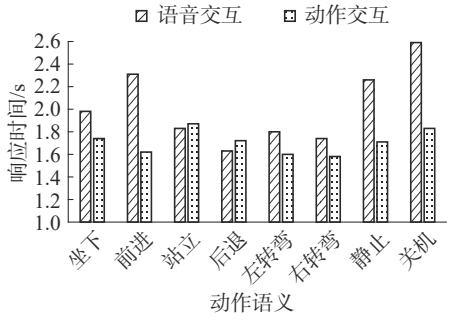


图 7　语音与动作交互的响应时间比较

Fig. 7　Comparison between the response time of speech and action interaction

通过响应时间和准确率的实验对比, 得出动作交互相比于语音交互更具有优越性, 不受环境噪声的影响, 空间距离影响较小, 适应环境能力强, 保证了实时性的同时, 满足在交互过程中的快速响应, 且准确率高。

4　结束语

针对运动信息和骨架关节信息独立建模不能表达充分的问题, 本文为了融合时空域的运动特征, 提出了基于姿态运动时空域融合的图卷积模型 (PM-STFGCN)。使用基于姿态运动的时空域融合模块 (PM-STF), 融合时域和空域特征并自适应增强特征。为了抑制时域上的干扰并学习运动姿态的表征, 本文定义了一种基于局部姿态运动的时域关注度模块 (LPM-TAM)。在大型骨架数据集上进行对比实验, 与目前性能较好的方法进行比较, 本文的方法更具有竞争性, 设计的动作交互系统在准确率和响应时间上都优于语音识

别，可以适应复杂的背景条件，体验者的交互效率更高。

参考文献：

[1] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 568−576.

[2] BAGAUTDINOV T, ALAHI A, FLEURET F, et al. Social scene understanding: end-to-end multi-person action localization and collective activity recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3425−3434.

[3] WANG Heng, SCHMID C. Action recognition with improved trajectories[C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 3551−3558.

[4] CAO Zhe, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1302−1310.

[5] CHEN Yilun, WANG Zhicheng, PENG Yuxiang, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7103−7112.

[6] 龚冬颖, 黄敏, 张洪博, 等. RGBD 人体行为识别中的自适应特征选择方法 [J]. 智能系统学报, 2017, 12(1): 1–7.
GONG Dongying, HUANG Min, ZHANG Hongbo, et al. Adaptive feature selection method for action recognition of human body in RGBD data[J]. CAAI transactions on intelligent systems, 2017, 12(1): 1–7.

[7] 姬晓飞, 王昌汇, 王扬扬. 分层结构的双人交互行为识别方法 [J]. 智能系统学报, 2015, 10(6): 893–900.
JI Xiaofei, WANG Changhui, WANG Yangyang. Human interaction behavior-recognition method based on hierarchical structure[J]. CAAI transactions on intelligent systems, 2015, 10(6): 893–900.

[8] 庄伟源, 成运, 林贤明, 等. 关键肢体角度直方图的行为识别 [J]. 智能系统学报, 2015, 10(1): 20–26.
ZHUANG Weiyuan, CHENG Yun, LIN Xianming, et al. Action recognition based on the angle histogram of key parts[J]. CAAI transactions on intelligent systems, 2015, 10(1): 20–26.

[9] 徐志通, 骆炎民, 柳培忠. 联合加权重构轨迹与直方图熵的异常行为检测 [J]. 智能系统学报, 2018, 13(6): 1015–1026.
XU Zhitong, LUO Yanmin, LIU Peizhong. Abnormal behavior detection of joint weighted reconstruction trajectory and histogram entropy[J]. CAAI transactions on intelligent systems, 2018, 13(6): 1015–1026.

[10] 吴云鹏, 赵晨阳, 时增林, 等. 基于流密度的多重交互集体行为识别算法 [J]. 计算机学报, 2017, 40(11): 2519–2532.
WU Yunpeng, ZHAO Chenyang, SHI Zenglin, et al. A flow density based algorithm for detecting coherent motion with multiple interaction[J]. Chinese journal of computers, 2017, 40(11): 2519–2532.

[11] 陈婷婷, 阮秋琦, 安高云. 视频中人体行为的慢特征提取算法 [J]. 智能系统学报, 2015, 10(3): 381–386.
CHEN Tingting, RUAN Qiuqi, AN Gaoyun. Slow feature extraction algorithm of human actions in video[J]. CAAI transactions on intelligent systems, 2015, 10(3): 381–386.

[12] 丁重阳, 刘凯, 李光, 等. 基于时空权重姿态运动特征的人体骨架行为识别研究 [J]. 计算机学报, 2020, 43(1): 29–40.
DING Chongyang, LIU Kai, LI Guang, et al. Spatio-temporal weighted posture motion features for human skeleton action recognition research[J]. Chinese journal of computers, 2020, 43(1): 29–40.

[13] 莫宏伟, 汪海波. 基于 Faster R-CNN 的人体行为检测研究 [J]. 智能系统学报, 2018, 13(6): 967–973.
MO Hongwei, WANG Haibo. Research on human behavior detection based on Faster R-CNN[J]. CAAI transactions on intelligent systems, 2018, 13(6): 967–973.

[14] 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, DOI: 10.11992/tis.201812029.
JI Xiaofei, XIE Xuan, Ren Yan. Human interaction recognition and prediction algorithm based on Deep Learning [J]. CAAI transactions on intelligent systems, DOI: 10.11992/tis. 201812029.

[15] 谢昭, 周义, 吴克伟, 等. 基于时空关注度 LSTM 的行为识别 [J/OL]. 计算机学报: (2019-12-17) http://kns.cnki.net/kcms/detail/11.1826.TP.20191227.1658.002.html.
XIE Zhao, ZHOU Yi, WU Kewei, et al. Activity recognition based on spatial-temporal attention LSTM[J/OL] Chinese journal of computers: (2019-12-17) http://kns.cnki.net/kcms/detail/11.1826.TP.20191227.1658.002.html.

[16] 王传旭, 胡小悦, 孟唯佳, 等. 基于多流架构与长短时记忆网络的组群行为识别方法研究 [J]. 电子学报, 2020, 48(4): 800–807.
WANG Chuanxu, HU Xiaoyue, MENG Weijia, et al. Research on group behavior recognition method based on multi-stream architecture and long short-term memory network[J]. Acta electronica sinica, 2020, 48(4): 800–807.

[17] 郑兴华, 孙喜庆, 吕嘉欣, 等. 基于深度学习和智能规划的行为识别 [J]. 电子学报, 2019, 47(8): 1661–1668.
ZHENG Xinghua, SUN Xiqing, LU Jiaxin, et al. Action recognition based on deep learning and artificial intelligence planning[J]. Acta electronica sinica, 2019, 47(8): 1661–1668.

[18] 张冰冰,葛疏雨,王旗龙,等.基于多阶信息融合的行为识别方法研究 [J/OL]. 自动化学报, [2020-06-17] DOI: 10.16383/j.aas.c180265.
ZHANG Bingbing, GE Shuyu, WANG Qilong, et al. Multi-order Information Fusion Method for Human Action Recognition[J/OL]. ACTA automatica sinica, [2020-06-17] DOI: 10.16383/j.aas.c180265.

[19] LIU Jun, SHAHROUDY A, XU Dong, et al. Spatio-temporal LSTM with trust gates for 3d human action recognition[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 816−833.

[20] LI Chao, ZHONG Qiaoyong, XIE Di, et al. Skeleton-based action recognition with convolutional neural networks[C]//Proceedings of 2017 IEEE International Conference on Multimedia and Expo Workshops. Hong Kong, China, 2017: 597−600.

[21] YAN Sijie, XIONG Yuanjun, LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 7444−7452.

[22] SHI L, ZHANG Y, CHENG J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J/OL]. [2020-06-01] https://arxiv.org/abs/1912.06971, 2019.

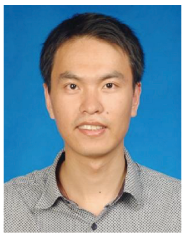
[23] LIU Ziyu, ZHANG Hongwen, CHEN Zhenghao, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 143−152.

[24] PENG W, HONG X, CHEN H, et al. Learning Graph Convolutional Network for Skeleton-based Human Action Recognition by Neural Searching[C]//Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence. New York, USA, 2020: 2669−2676.

[25] OBINATA Y, YAMAMOTO T. Temporal extension module for skeleton-based action recognition[J/OL]. [2020-03-19] http://arxiv.org/abs/2003.08951.

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

作者简介：



钟秋波，副教授，博士，宁波工程学院机器人学院执行副院长，主要研究方向为机器人智能控制、计算机视觉图像处理、机器人运动控制。先后主持和参与横、纵向科研项目 20 多项。发表学术论文 20 余篇。

郑彩明，硕士研究生，主要研究方向为机器人智能控制、计算机视觉、图像处理、机器人运动控制。

朴松昊，教授，博士生导师，中国人工智能学会常务理事，机器人文化艺术专业委员会主任，主要研究方向为机器人环境感知与导航、机器人运动规划、多智能体机器人协作。主持或参加国家自然科学基金、国家“863”计划重点、教育部“985”等多个项目。发表学术论文 60 余篇。