DOI: 10.19650/j.cnki.cjsi.J2310939

基于时空张量融合的人体骨架行为自适应识别方法*

建中华[1],南 静[2],刘 鑫[1],代 伟[1,2]
(1. 中国矿业大学人工智能研究院 徐州 221116; 2. 中国矿业大学信息与控制工程学院 徐州 221116)

摘 要:针对人体行为的空间复杂性和时间差异性问题,提出了一种基于时空张量融合的人体骨架行为自适应识别方法。首先充分利用人体行为骨架序列的帧内空间关系和帧间时间关系,构建相邻帧时空特征张量;其次通过计算相邻帧时空特征张量的差异性获取关键相邻帧时空特征张量并组成行为时空特征张量;之后利用行为时空特征张量的空间特征差异和多尺度时间卷积构建行为时空特征张量自适应注意力机制,完成行为时空特征融合;最后,使用深度随机配置网络根据行为时空特征融合张量识别人体行为。使用 NTU RGB-D 数据集进行实验仿真,识别准确率达到 84.57%,并且设计相应的系统进行实际应用验证,结果表明本文所提方法是一种适合应对人体行为空间复杂性和时间差异性问题的人体行为识别方法。

关键词: 人体行为识别;人体骨架;注意力机制;关键帧;时空特征
中图分类号: TP391.4 TH164 **文献标识码**: A **国家标准学科分类代码**: 120.3 520.2

Adaptive recognition method of human skeleton action with spatial-temporal tensor fusion

Jian Zhonghua[1], Nan Jing[2], Liu Xin[1], Dai Wei[1,2]
(*1. Artificial Intelligence Research Institute, China University of Mining and Technology, Xuzhou 221116, China;*
2. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: To address space complexity and time difference of human action, an adaptive recognition method of human skeleton action with spatial-temporal tensor fusion is proposed. Firstly, the spatial-temporal feature tensors of adjacent frames are established by making full use of the intra-frame spatial relationship and the inter-frame temporal relationship of the human action skeleton sequences. Secondly, the difference of spatial-temporal feature tensors of adjacent frames is calculated to achieve the spatial-temporal feature tensors of key adjacent frames and compose the behavior spatial-temporal feature tensors. Then, the spatial feature difference of the action spatial-temporal feature tensors and the multi-scale temporal convolution is used to construct the adaptive attention mechanism of the behavior spatio-temporal feature tensors to complete the fusion of action spatial-temporal features. Finally, a deep stochastic configuration network is used to recognize human action according to the spatial-temporal feature fusion tensor of action. The NTU RGB-D data set was used for experimental simulation, and the recognition accuracy reached 84.57%. The corresponding system is designed for practical application verification. The results show that the proposed method is suitable for dealing with the space complexity and time difference of human action recognition.

Keywords: human action recognition; human skeleton; attention mechanism; keyframe; spatial-temporal feature

0 引 言

人体行为识别(human action recognition, HAR)是指基于收集的人体多源信息,构建多元信息与行为特征映射关系,使计算机或其他智能体理解和识别人体行为。HAR 在诸多领域已经展现出巨大潜力,如在矿场作业生产中监控并识别现场作业人员的行为,能够帮助作业人员规范其行为,从而降低人为事故发生率,保证生产安全运行[1];在汽车自动驾驶过程中识别环境中的人体行为,有助于制定行驶策略,避免碰撞,保证道路安全[2];在智能机器人行业中帮助机器人精准定位和识别人体行为,能够使机器人更加安全高效地作业[3]。

收稿日期:2023-01-01 Received Date: 2023-01-01
*基金项目:国家自然科学基金面上项目(61973306)、江苏省优秀青年基金(BK20200086)、江苏省研究生科研与实践创新计划项目(KYCX22_2558)资助

人体行为是具有空间复杂性和时间差异性的动态过程[4],空间复杂性是指在不同场景中不同行为人做不同动作的姿态差异,时间差异性是指人体行为发生时刻和持续时间不可预测。目前,按照输入数据形式的不同可以分为基于可穿戴传感器的 HAR 和基于视频的 HAR 两大类[5]。前者主要依靠可穿戴感器收集人体各部位运动的速度、加速度等数据,进一步提取特征和识别人体行为。本文前期工作利用智能手机提取了人体行为的时域和频域特征,进一步提出基于流形正则化和 QR 分解的轻量化人体行为识别方法[6]。但此类方法依赖可穿戴传感器,传感器类型和位置等差异导致难以建立通用的 HAR 模型。基于视频的 HAR 以 RGB 视频流数据为基础[7-8],具有相对的数据稳定性,有利于建立普适的识别模型,减轻了对设备的依赖,但其存在易受光照和背景等因素影响的弊端。为此诸多学者着力于面向视频的 HAR 方法的优化[9-10]。提取视频流数据中人体的 2D 或 3D 骨架节点坐标信息,以此进行行为识别已经成为当前一大主流方法[11],这是由于人体骨架节点数据具有噪声小、鲁棒性强和避免视频背景干扰的优点[12]。此外,OpenPose、Kinect 等人体骨架节点信息获取软硬件工具的发展,为基于骨架的 HAR 方法研究提供了便利[13-14]。

基于骨架的 HAR 主要分为特征表征和动作识别[15],其中动作识别多采用支持向量机(support vector machines, SVM)等分类器来实现,众多的研究成果更关注行为特征的表征。文献[16]计算人体行为相邻骨架帧之间的协方差,构建协方差矩阵表示人体行为时空特征。文献[17]根据人体骨架坐标计算空间特征后,使用聚类算法筛选关键骨架帧,将关键骨架帧空间特征的集合作为时间特征。文献[18]使用流形表示行为的时空特征,一个流形由多个行为帧组成,根据待测序列流形和类别流形模板的最近距离识别行为。文献[19]对人体骨架节点的坐标滤波后直接使用坐标作为空间特征,使用骨架时间序列的组合作为行为的时间特征后,利用模板匹配的方法为骨架序列分类行为。文献[20]使用帧内人体各关节之间的距离作为空间特征,多帧空间特征的组合作为时间特征,计算待测序列与模板序列的距离识别行为,在计算两序列距离时加入了动态时间规整(dynamic time warping, DTW)来缓解两匹配序列长度不一致的问题。文献[21]根据序列间的距离为每类行为确定符号模板,根据待测骨架帧到每个符号的距离构建概率矩阵,组合概率最高的符号序列,与符号模板匹配识别行为。文献[22]基于动力学使用骨架节点帧内相对位置表示空间特征,节点帧间速度表示时空特征,采用双线性分类器迭代计算行为特征的权重,并引入 DTW 和时间金字塔优化行为时序表示。文献[23]计算人体关节帧内相对位置和角度作为空间特征,计算人体关节帧间速度作为时间特征,利用主成分分析压缩特征维度。文献[24]将人体行为骨架序列表示为关节运动轨迹,提取轨迹图的视觉特征。上述方法试图利用人体的几何特征或行为动力学的行为时空特征表示人体行为,获得了比关节点更好的特性表征,但是表征能力仍然有限且难以选取有效的特征[25]。当前,诸多学者使用深度学习探索行为特征的自主表征[26],其中卷积神经网络(convolutional neural networks, CNN)和长短期记忆网络(long short-term memory, LSTM)由于在特征学习和时序建模方面的优异表现被应用于 HAR 研究。如文献[27]利用 CNN 从人体骨架序列中提取低维和高维的空间特征,使用 LSTM 基于空间特征提取时间特征,通过全连接网络(full connect neural network, FCN)识别行为。文献[28]将人体骨架信息作为空间特征和视频帧信息融合并使用 LSTM 提取时间特征后,使用 FCN 识别行为。图卷积网络(graph convolutional network, GCN)通过提取人体骨架拓扑结构的特征[29]增强了特性表征,如,文献[30]以有向图表示人体骨架,将分类残差卷积引入 GCN 中,保证了网络挖掘特征的多样性。文献[31]在使用 GCN 提取人体行为空间特征后,进一步引入扩张因果卷积和残差结构以构建时域扩张残差网络,从而实现行为识别。文献[32]提出时空图卷积网络(spatial temporal GCN, ST-GCN),使用多层 GCN 和基于一维卷积的时间卷积(temporal convolutional network, TCN)[33]提取行为时空特征,进而利用 FCN 识别行为。文献[34]在 ST-GCN 的基础上进一步考虑骨架不相邻节点关系,改善了网络性能。为了使网络更关注重要特征,基于注意力机制的 Transformer 模型被引入到 HAR 研究中[35]。文献[36]使用 Tranformer 模型,利用编码、解码的方式识别行为,并通过增减网络注意力头数和层数,设计不同复杂度的模型。文献[37]将 ST-GCN 与 Transformer 注意力网络结合,使用 ST-GCN 提取低维的时空特征,进而基于双流网络实现帧内特征注意和帧间特征注意。

上述以深度学习为基础的人体行为识别模型获得了较高的准确率,但值得注意的是,其模型结构和计算复杂度较高,网络训练相对困难,对硬件要求较高[12,38]。此外,现有 HAR 研究工作大多根据人为划分的持续数秒的行为序列流数据识别行为,模型缺乏人体行为发生时刻和持续时间的自主判别能力,对于人体行为时间差异性的应对能力较弱,欠缺实时性[39]。

1　STAR

针对上述人体骨架行为识别现存问题,本文提出基于时空张量融合的自适应识别方法(spatial-temporal tensor fusion for adaptive recognition, STAR),其结构如图 1 所示。首先相邻帧时空特征构建模块根据骨架帧

的时空关系，构建相邻帧时空特征张量增强数据表达性。其次，关键相邻帧时空特征张量注意模块通过计算相邻帧时空特征张量的差异性保留关键相邻帧时空特征张量。然后在行为时空特征张量空间注意模块和行为时空特征张量时间注意模块实现行为时空特征张量的时空注意和融合，最后使用深度随机配置网络（deep stochastic configuration networks，DeepSCNs）识别行为。本文通过对 NTU RGB-D 数据集重构，模拟行为发生时刻和持续时间不确定情况下的 HAR 并进行实验仿真，结果表明所提出的模型是一种适合应对人体行为空间复杂性和时间差异性问题的人体行为识别方法。

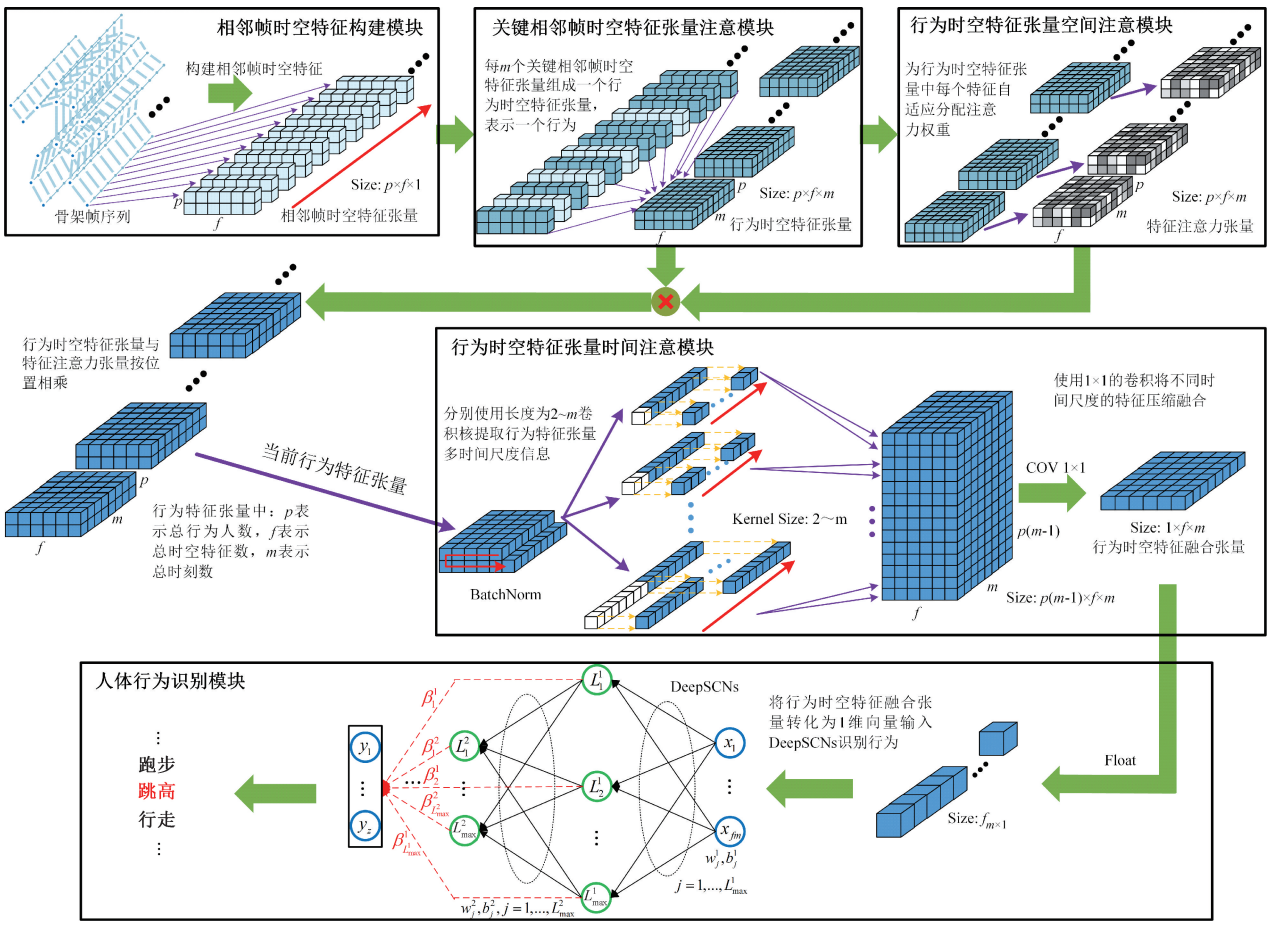


图 1 STAR 结构图

Fig. 1 Architecture diagram of STAR

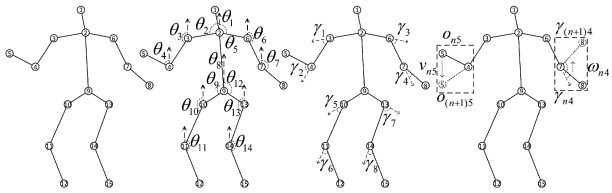
1.1 相邻帧时空特征构建模块

本文利用人体行为帧内空间关系和帧间时间关系构建相邻帧时空特征张量，增强数据的表征能力。由于人体行为主要反应在各肢体的角度、各肢体之间的角度、各部位速度和各角速度的变化，对于骨架序列的一帧计算帧内肢体角度 θ，帧内关键肢体角度 γ 表示骨架帧内空间特征，与上一帧的帧间骨架速度 v 和帧间角速度 ω 表示相邻骨架帧的时间特征。

具体来说，如图 2(a)所示，选取 15 个人体骨架节点来表示人体各部分，根据人体生理结构，15 个人体骨架节点连接为 14 段可以表示人体肢体的刚体向量。对于一段长度为 N 的人体骨架行为序列，使用集合 $\boldsymbol{O}=\{o_{ni}\mid n=1,\cdots,N,\ i=1,\cdots,15\}$ 来表示序列中的所有的骨架节点，使用集合 $\boldsymbol{L}=\{l_{ni}\mid n=1,\cdots,N,\ i=1,\cdots,14\}$ 来表示序列中所有的骨架肢体，令 $(o_{nix},o_{niy},o_{niz})$ 表示第 n 帧中第 i 个骨架节点的三维坐标，$(l_{nix},l_{niy},l_{niz})$ 表示第 n 帧中第 i 段人体肢体的三维坐标。

如图 2(b)所示，帧内肢体角度是指每帧内 14 段人体肢体向量与三维坐标轴 Z 轴正方向所成夹角。令 Z 轴正方向为(0, 0, 1)，令 θ_{ni} 表示第 n 帧中第 i 个帧内肢体角度，则其计算公式如式(1)所示。

$$\theta_{ni}=\arccos\left(\frac{l_{niz}}{\sqrt{(l_{nix}^2+l_{niy}^2+l_{niz}^2)}}\right),\ i=1,\cdots,14 \quad (1)$$



(a) 选取的15个骨架节点 (a) The 15 selected skeleton nodes　(b) 帧内肢体角度 (b) The limb angles of intra-frame　(c) 帧内关键肢体角度 (c) The key limb angles of intra-frame　(d) 帧间骨架速度和角速度 (d) The velocity and angular velocity of skeleton of intra-frame

图 2　相邻骨架帧时空特征示意图

Fig. 2　Diagram of spatial-temporal features of adjacent skeleton frame

如图 2(c)所示,帧内关键肢体角度是指每帧内人体骨架肢体向量所成的 8 个关键夹角。令 $\gamma_{n(i,j)}$ 表示第 n 帧中第 i 段肢体和第 j 段人体肢体组成的帧内关键肢体角度,则其计算公式如式(2)所示。

$$\gamma_{n(i,j)} = \arccos\left(\frac{l_{nix}l_{njx} + l_{niy}l_{njy} + l_{niz}l_{njz}}{\sqrt{(l_{nix}^2 + l_{niy}^2 + l_{niz}^2) \times (l_{njx}^2 + l_{njy}^2 + l_{njz}^2)}}\right),$$
$$i,j = 1,\cdots,14 \tag{2}$$

如图 2(d)左侧所示,帧间骨架速度是指同一人体骨架节点在相邻两帧之间的运动速度。令 $(v_{nix},v_{niy},v_{niz})$ 表示第 n 帧中第 i 个骨架节点的三维速度,由于无法计算第 1 帧与前 1 帧的帧间骨架速度,故当 $n=1$ 时,令 $(v_{1ix},v_{1iy},v_{1iz})=(0,0,0)$,$(i=1,\cdots,15)$。则其计算公式如式(3) 所示,其中 t 表示相邻两帧间隔时间。

$$(v_{nix},v_{niy},v_{niz}) = \frac{(o_{nix},o_{niy},o_{niz}) - (o_{(n-1)ix},o_{(n-1)iy},o_{(n-1)iz})}{t},$$
$$n = 2,\cdots,N;\ i = 1,\cdots,15 \tag{3}$$

如图 2(d)右侧所示,帧间肢体角速度是指同一肢体关键角度在相邻两帧之间的角速度。令 ω_{ni} 表示第 n 帧第 i 个帧间肢体角速度,由于无法计算第 1 帧与前 1 帧的帧间肢体角速度,故当 $n=1$ 时,令 $\omega_{1i}=0$,$(i=1,\cdots,8)$,则其计算公式如式(4) 所示。

$$\omega_{ni} = \frac{\theta_{ni} - \theta_{(n-1)i}}{t},\ n = 2,\cdots,N,i = 1,\cdots,8 \tag{4}$$

综上所述,使用 75 维时空特征表示相邻骨架帧的时空特征,其中空间特征包括 14 维帧内肢体角度特征 θ 和 8 维帧内关键肢体角度特征 γ,时间特征包括 45 维相邻骨架帧间骨架速度特征 v 和 8 维相邻骨架帧间肢体角速度特征 ω。为简单起见,本文只对骨架帧中置信度最高的两人计算时空特征。故如图 1 中相邻骨架帧时空特征构建模块所示,每帧中的两个人体骨架时空特征可以形成 2×75×1 的相邻骨架帧时空特征张量。

1.2　关键相邻帧时空特征张量注意模块

众所周知,对于一个行为骨架序列,其中只有部分包含重要行为信息的关键骨架帧[5],筛选出关键骨架帧利于识别行为。本文使用相邻帧时空特征张量表示行为骨架序列中某一帧的帧内空间关系和该帧与前一帧的时间关系,并将其视为空间向量,利用余弦相似度判断相邻骨架帧时空特征张量间的差异性,自适应的获取关键相邻帧时空特征张量,组成能够表示一个人体行为的行为时空特征张量。

设余弦相似度阈值为 T,m 个关键相邻帧时空特征张量的组合可以表示一个行为,本文通过实验验证确定其数值。令 $F=\{f_n \mid n=1,\cdots,N\}$ 表示一段包含 N 个相邻帧时空特征张量序列,f_i 表示序列 F 中第 i 个相邻帧时空特征张量,当前关键相邻帧时空特征张量为 c,FT 表示包含 m 个关键相邻帧时空特征张量的当前行为时空特征张量。具体步骤如下:

针对一段行为骨架序列的相邻帧时空特征张量 F,首先令 F 中的 f_1 为 c 并加入 FT 中,之后根据式(5)计算 f_2 与 c 的余弦相似度 S。

$$S = \frac{\sum_{i=1}^{75}(f_{ni} \times c_i)}{\sqrt{\sum_{i=1}^{75} f_{ni}^2 \times \sum_{i=1}^{75} c_i^2}},\ n = 1,\cdots,N \tag{5}$$

若 $S>T$,则 f_2 为非关键骨架帧张量,放弃 f_2,继续 f_3 的筛选;若 $S<T$,则 f_2 为关键骨架帧张量,将 f_2 设置为 c,若 FT 中关键相邻帧时空特征张量个数未达到 m,将 f_2 加入 FT 中,反之将 f_2 加入到下一 FT 中。继续进行 f_3 的筛选,直到完成序列 F 中所有相邻帧时空特征张量的筛选,舍弃末尾关键相邻帧时空特征张量个数不足 m 个的行为时空特征张量。

1.3　行为时空特征张量空间注意模块

由于人体在做动作时身体的各部分并非同步运动,不同行为不同部位的移动程度不同,区分度也不同,应当为区分度高的行为特征自适应的分配不同的注意力权重。本文基于行为时空特征张量中行为特征的差异性来为张量中每个位置的行为特征计算注意力权重。一个行为时空特征张量自适应注意力权重分配准则如下:

1)同类型的行为特征注意力权重和为 1,每个行为特征张量中所有的注意力权重和为 4;

2)不同行为特征需要自适应分配不同的权重;

3)变化幅度越大的特征越关键,需要分配更多的注意力权重。

如图 1 中行为时空特征张量空间注意模块所示,令某个行为时空特征张量 $FT=\{ft_{ijk} \mid i=1,2,j=1,\cdots,75,k=1,\cdots,m\}$,令其对应的特征注意力张量 $A=$

$\{a_{ijk} \mid i=1, 2, j=1, \cdots, 75, k=1, \cdots, m\}$。其中 i 表示第 i 个行为人，j 表示第 j 个行为特征，k 表示第 k 个时刻。由于 FT 包含 m 个关键相邻帧时空特征张量，故共有 m 个时刻。首先，分别计算每个行为人每个特征在所有时刻 m 上的方差 ξ_{ij}，之后根据 ξ_{ij} 在同类型行为特征总方差的比重为每个行为人的每个特征 ft_{ij} 分配注意力系数 a_{ij}，计算公式如式(6)所示，其中 W 表示同类型行为特征的总数。

$$a_{ij} = \frac{\xi_{ij}}{\sum_{i,j=1}^{i=2,j=W} \xi_{ij}} \tag{6}$$

根据式(7)计算 ft_{ij} 在所有时刻上的平均值 vft_{ij}。

$$vft_{ij} = \frac{\sum_{k=1}^{m} ft_{ijk}}{m};\ i = 1,2, j = 1,2,\cdots,75 \tag{7}$$

进而根据式(8)计算 ft_{ikj} 与 vft_{ij} 的距离 d_{ijk}。

$$d_{ijk} = |ft_{ijk} - vft_{ij}|, i = 1,2; j = 1,2,\cdots,75; k = 1,\cdots,m \tag{8}$$

最后根据 d_{ijk} 为行为时空特征张量中每个特征 ft_{ijk} 分配注意力权重 a_{ijk}，计算公式如式(9)所示。

$$a_{ijk} = \frac{d_{ijk}}{\sum_{k=1}^{m} d_{ijk}} \times a_{ij}, i = 1,2, j = 1,2,\cdots,75 \tag{9}$$

通过计算特征差异性的方式为行为时空特征张量中的每个特征自适应分配注意力权重，物理意义明确且计算量小。将行为时空特征张量和特征注意力张量对应位置相乘完成行为时空特征的空间注意，之后输入时间注意机制模块。

1.4 行为时空特征张量时间注意模块

为了保证模型的轻量化，本文采用多尺度时间卷积针对行为时空特征张量空间注意模块的输出，构建时间注意机制，完成行为时空特征的融合。

如图 1 中行为时空特征张量时间注意模块所示，其针对行为时空特征张量，做批归一化处理，之后在行为时空特征张量中的每个特征的所有时刻上按图示次序和方向做时间卷积，由于行为特征张量中包含 m 个关键相邻帧的时空特征信息，故针对不同时间尺度设置时间卷积核的长度为 $2\sim m$，保证模型对人体行为中不同时间尺度的关注。为了保持卷积输入输出时间尺度对应，在行为特征张量数据前补 0。令某个行为时空特征张量 $FT = \{ft_{ijk} \mid i = 1, 2, j = 1, \cdots, 75, k = 1, \cdots, m\}$，时间卷积输出的行为时空特征张量 $OUTFT = \{outft_{ijk} \mid i = 1, 2, j = 1, \cdots, 75, k = 1, \cdots, m\}$，其中 i 表示第 i 个行为人，j 表示第 j 个行为特征，k 表示第 k 个时刻，则对于行为时空特征张量中特征 f_{ij} 的所有时刻，如式(10)所示进行时间卷积，其中 ζ 表示激活函数，s 表示卷积核长度，$padding$ 表示在第 1 时刻前补零，s-1 表示补零数，w 和 b 表示权值和偏置。

$$outft_{ij} = \zeta\left(\sum_{s=2}^{m} (w(ft_{ij} + padding(s-1))) + b\right),$$
$$i = 1,2; j = 1,\cdots,75 \tag{10}$$

为了约简特征维数，使用 1×1 的卷积，将 $m-1$ 个时间尺度的行为时空特征张量压缩融合为 1 个 $1\times f\times m$ 的行为时空特征融合张量。

1.5 人体行为识别模块

随机配置网络(stochastic configuration networks, SCNs)作为一种增量式的单隐层前馈神经网络模型，在网络构建过程中，通过基于残差构建的监督机制，在候选参数中为新增节点自适应选择最优参数，从而保证模型的轻量性和无限逼近性[40]。由于 SCNs 网络具有良好的性能，被广泛应用[41-43]。然而单一隐含层映射能力有限。为提高模型识别性能，本文基于 DeepSCNs[44] 建立了行为识别模块。如图 3 所示，首先基于残差监督约束，增量式构建第 1 层隐层，直到满足预设最大节点数停止；然后，将第 1 层隐层的输出作为输入传输至第 2 层隐层，并使用相同的构建方式构建第 2 层隐含层，重复上述步骤直到满足预设的网络残差要求或达到网络最大隐含层数为止。

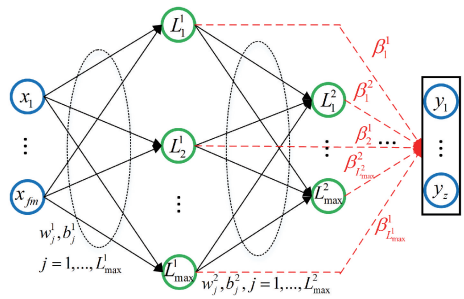


图 3 DeepSCNs 的网络模型结构图

Fig. 3 Network architecture of DeepSCNs

对于给定目标输出 $Y = \{y_1, y_2, \cdots, y_N\}$，$y_i = [y_{i,1}, y_{i,2}, \cdots, y_{i,z}]$：$R^d \rightarrow R^z$($z$ 表示人体行为类别总数)、网络输入 $X = \{x_1, x_2, \cdots, x_N\}$，$x_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,fm}]$：$R^d \rightarrow R^{fm}$($fm$ 表示行为时空特征维数)、网络最大隐层数 M、第 i 层隐层最大节点数 $L_{\max}^i$、允许残差 ε、初始残差 $\varepsilon_0 = \boldsymbol{Y}^{\mathrm{T}}$、激活函数 $\zeta(0 < \|\zeta\| < k)$、参数随机生成区间 $[-\lambda, \lambda]$、每次随机配置参数的组数 g、$r \in (0, 1)$ 和非负递减数列 $\{\eta_J\}$ $(0 < \eta_J < (1-r))$，DeepSCNs 的具体构建过程如下：

第 1 层隐层随机构建方法：首先在区间 $[-\lambda, \lambda]$ 中为第 1 层第 1 个节点随机配置 g 组权值 $w_1^1 = [w_{1,1}^1, w_{1,2}^1, \cdots, w_{1,fm}^1]$ 和偏置 b_1^1，然后根据隐层节点参数选取方法从中自适应选取最优参数。进而根据输出权值和网络

输出构建方法计算网络输出权值 β^* 和网络输出 $F_J(X)$。最后根据网络残差计算方法计算网络当前残差 $\boldsymbol{\varepsilon}_J$,若 $\|\boldsymbol{\varepsilon}_J\|_F \leqslant \varepsilon$,则网络构建完毕;若 $\|\boldsymbol{\varepsilon}_J\|_F > \varepsilon$ 且未达到第 1 层最大节点数,则重复上述步骤,继续向第 1 层添加节点,增加网络的特征映射能力。直到第 1 层中节点数达到 $L^1_{\max}$,若 $\|\boldsymbol{\varepsilon}_J\|_F$ 仍不满足要求,证明仅增加隐层中的节点难以满足目标任务需求,进而开始下一隐层的构建。

第 n 层隐层随机构建方法:第 n 层隐层以第 $n-1$ 层的隐层输出为本层输入,类似于第 1 层隐层的构建方法,首先随机配置 g 组第 n 层第 1 个节点的权值 $w^n_1 = [w^n_{1,1}, w^n_{1,2}, \cdots, w^n_{1,L^{n-1}_{\max}}]$ 和偏置 b^n_1,之后选取最优参数并更新 β^*、$F_J(X)$ 和 ε_J。根据 $\|\boldsymbol{\varepsilon}_J\|_F$ 决定继续向本层中添加节点或开始构建下一隐层。重复上述步骤,直到满足残差要求或达到网络最大隐层数 M。

隐层节点参数选取方法:选取满足式(11)且使得 $\sum_{q=1}^{z} \langle \boldsymbol{\varepsilon}_{J,q}, \boldsymbol{\phi}^i_j \rangle^2$ 值最大的一组参数为最优参数。

$$\begin{cases} \langle \boldsymbol{\varepsilon}_{J,q}, \boldsymbol{\phi}^i_j \rangle^2 \geqslant (1-r-\eta_J) k^2 \|\boldsymbol{\varepsilon}_{J,q}\|^2 \\ \boldsymbol{\phi}^i_j = (\zeta((\boldsymbol{w}^i_j)^{\mathrm{T}}(\boldsymbol{\Phi}^{i-1})^{\mathrm{T}} + b^i_j))^{\mathrm{T}} \\ \boldsymbol{\Phi}^{i-1} = [\phi^{i-1}_1, \phi^{i-1}_2, \cdots, \phi^{i-1}_{L^{i-1}_{\max}}] \end{cases},$$

$$q = 1,2,\cdots,z \tag{11}$$

其中,J 表示当前网络中所有节点的总数,$\boldsymbol{\varepsilon}_J$ 表示网络当前残差,$\langle .,. \rangle$ 表示两向量的内积,$\boldsymbol{\phi}^i_j$ 表示新增节点输出,$\boldsymbol{\Phi}^{i-1}$ 表示第 $i-1$ 层隐层输出,令 $i=1$ 时,$\boldsymbol{\Phi}^0 = [x_1, x_2, \cdots, x_N]^{\mathrm{T}}$,$\boldsymbol{w}^i_j$、$b^i_j$ 表示节点 L^i_j 的候选参数。

输出权值及网络输出构建方法:根据式(12)通过最小二乘法构建网络的输出权值。根据式(13)构建网络输出。其中 H 为网络总隐层数,L_i 为第 i 层隐层中节点个数,$F_J(X)$ 表示当前网络的输出。

$$\beta^* = \mathrm{argmin}_{\beta} \left\| Y - \sum_{i=1}^{H} \sum_{j=1}^{L_i} \beta^i_j \boldsymbol{\phi}^i_j \right\|^2 \tag{12}$$

$$\boldsymbol{F}_J(X) = \sum_{i=1}^{H} \sum_{j=1}^{L_i} \beta^i_j \boldsymbol{\phi}^i_j \tag{13}$$

网络残差计算方法,根据式(14)计算网络当前残差。

$$\boldsymbol{\varepsilon}_J = \boldsymbol{Y} - \boldsymbol{F}_J(X) = [\boldsymbol{\varepsilon}_{J,1}, \boldsymbol{\varepsilon}_{J,2}, \cdots, \boldsymbol{\varepsilon}_{J,q}],\ q = 1,2,\cdots,z \tag{14}$$

由于监督机制的约束,构建的 DeepSCNs 比 FCN 更为轻量化,由于每层隐含层和输出层之间的权值连接,DeepSCNs 比同节点数的 FCN 表达能力更强。并且文献[44]中通过数学分析证明了 DeepSCNs 的收敛性和无限逼近性。由于 DeepSCNs 的诸多优点,本文采用 DeepSCNs 作为本文网络模型行为识别模块的分类器。

2　实验研究

为验证本文方法的有效性,重构 NTU RGB-D 数据集[45]模拟人体行为发生频率和持续时间不确定的情况进行仿真实验,并与多种识别方法对比。

2.1　数据集介绍及评价指标

本文选择 NTU RGB-D 数据集作为实验初始数据集,该数据集是目前 HAR 领域中代表性较强的人体行为数据集。采集了 40 名年龄在 10~35 岁之间的志愿者,共计 60 个动作类的 56 000 个片段,由 3 个 Microsoft Kinect V2 相机同时捕获包含人体 25 个关节点的 3D 骨架数据。

评价标准是衡量网络模型鲁棒性和泛化性的主要方式。HAR 作为多分类问题,本文采用准确率(accuracy)作为评价指标对识别结果做出评价,准确率是指对于给定测试集,网络模型正确分类的样本数和总样本数的比值。

2.2　实验设置

本文仿真实验均是在 CPU 为 i7-11700K 3.6 GHz,RAM 为 16 GB,GPU 为 RTX3060 12 GB 的硬件平台上,采用 Pytorch 深度学习框架,利用 GPU 加速算法运行的情况下进行仿真实验。所有实验均选取数据集中的 80% 作为训练集,20% 作为测试集。

1)STAR 实验设置

本文所提 STAR 方法中相邻帧时空特征构建模块和行为时空特征张量空间注意模块无需确定实验参数,其余模块实验参数设置如下:

针对 NTU RGB-D 数据集中扔行为的某一段骨架序列,通过实验确定关键相邻帧时空特征张量注意模块中余弦相似度阈值 T 和能够表示一个行为的关键相邻帧时空特征张量个数 m。

设置不同余弦相似度阈值 T 选取关键相邻帧时空特征张量,使用选取的前 4 个关键相邻帧时空特征张量对应的骨架帧展示实验效果。结果如图 4 所示,$T=0.50$ 时,骨架帧变化较大,对行为过程体现效果差;$T=0.60$ 和 0.65 时骨架帧变化较小,包含冗余信息;$T=0.55$ 时选取的骨架帧变化合适,既能体现行为的过程,又没有冗余信息。故本文中选择 $T=0.55$ 的设置。

使用余弦相似度阈值 $T=0.55$ 选取关键相邻帧时空特征张量,展示不同关键相邻帧时空特征张量个数 m 对扔行为过程的体现效果。结果如图 5 所示,$m=6$ 时,不能体现完整的行为过程;$m=7$ 时,可以体现完整的行为过程;$m=8$ 时,能够体现完整行为过程且信息稍有冗余。本文为保证选取的行为时空特征张量对不同行为动态过

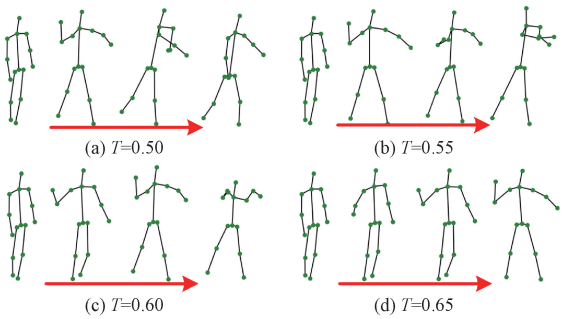


图 4 不同余弦相似度阈值选取效果图

Fig. 4 Diagram of different cosine similarity threshold selection effect

程表达的完整性,选择 $m=8$ 的设置。图 5(d)中展示了大多数方法中采用的随机抽取骨架帧的效果[27,37],与本文的方法相比其中包含过多冗余信息。

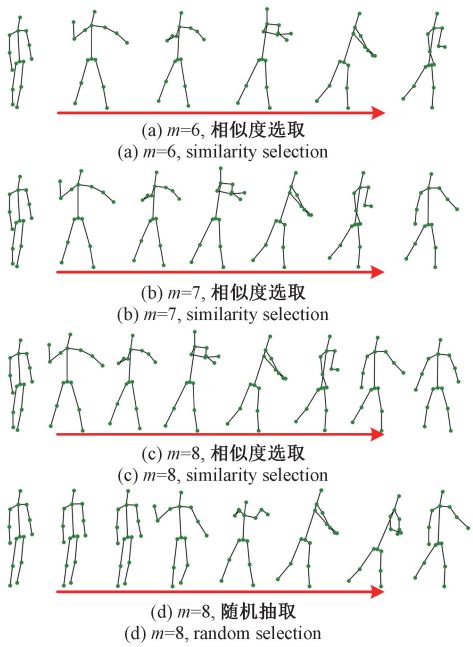


图 5 不同关键相邻帧时空特征张量个数表示行为效果图

Fig. 5 Diagram of using different key adjacent spatial-temporal feature tensors to represent action

行为时空特征张量时间注意模块中多尺度时间卷积网络中学习率和批量大小两个超参数较难确定,本文通过实验的方式根据网络分类准确率确定其数值,在试验不同参数时暂时使用包含 4 层隐层的 FCN 识别行为,4 层的节点数分别为 3 000、1 000、600、60。学习率和批量大小参数不同设置情况下网络分类准确率如图 6 所示。可以看到对于本数据集学习率设置为 0.000 5,批量大小设置为 32 最合适。因为学习率过大会使网络波动较大,难以找到最优解,过小会使网络收敛过慢,容易陷入局部最优解;批量过小会使每批中样本过少,信息不足,从而导致准确率不高,当批量大于 32 时训练损失减小的慢,训练相同轮数相同情况下准确率不高。实验中多尺度时间卷积的其他参数设置如下:损失函数为 Cross Entropy Loss,优化器为 Adam,训练轮数为 80,权值衰减为 0.001。

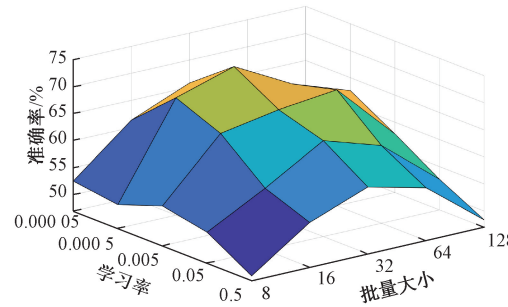


图 6 参数不同设置下网络准确率情况

Fig. 6 Accuracy with different parameter settings

本文根据经验设置人体行为识别模块中 DeepSCNs 的参数,相关参数设置如下:最大隐层数为 4,各隐层中最大节点数为 2 000,1 000,600,300,允许残差为 0.5,激活函数为 Sigmod,随机配置区间为[-1, 1],每次随机配置参数的组数为 100。

2)对比方法实验设置

选择多种现有算法与本文所提方法进行对比。由于传统的人体行为识别算法识别准确度普遍低于基于深度学习的方法[39],故传统方法中只选择算法性能较好的 DeepSCNs 进行比较,DeepSCNs 的参数与本文中人体行为识别模块使用相同设置。基于深度学习的方法虽然识别准确率高,但同时网络模型复杂,训练困难,对硬件要求较高,故本文主要选取模型复杂度较低且网络结构差异较大的方法提取行为特征,进而通过 3 层全连接网络识别行为,每层中的节点数为 1 600、800、60。GCN 能够处理人体空间拓扑结构信息,本文选择 GCN 作为一种对比方法,根据人体骨架拓扑结构定义邻接矩阵,第 1 层 GCN 设置 500 个节点,第 2 层 GCN 设置 200 个节点。由于人体行为具有时序特征,故本文选取 LSTM 作为另一种对比方法,第 1 层 LSTM 设置 800 个节点,第 2 层设置 200 个节点。由于人体行为是具有时空特性的动态过程,故本文选择文献[28]中的 CNN+LSTM 方法进行对比,第 1 层 CNN 设置 600 个节点,第 2 层 CNN 设置 300 个节点,第 1 层 LSTM 设置 800 个节点,第 2 层 LSTM 设置 200 个节点。上述基于深度学习方法的训练轮数均设置为 80 轮。

2.3 结果分析

使用本文所提方法中的相邻帧时空特征构建模块和关键相邻帧时空特征张量注意模块对 NTU RGB-D 数据集中每段骨架序列进行重构,进而通过行为时空特征张

量空间注意模块、时间注意模块和人体行为识别模块识别行为。

重构后的数据集共有 56 554 个样本,各类行为的具体分布如图 7 所示,各类行为样本数的分布大致均衡,最少类别样本数不高于 905,最多类别样本数不高于 954。其中前 40 类行为包含饮酒、吃饭、阅读等日常的单人行为,中间 9 类行为包含打喷嚏、摔倒等与健康相关的单人行为,最后 11 类行为包含拳击、踢腿、拥抱等多人行为。

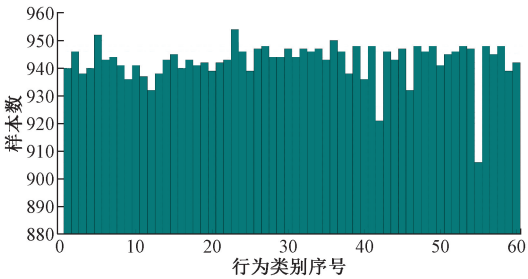


图 7　重构数据集中各类人体行为样本数分布

Fig. 7　The number distribution of various human action samples in the reconstructed data

各类行为识别准确率如图 8 所示。对于前 40 类日常单人行为的识别准确率大多在 80% ~ 90% 之间;第 2、19、23、24 和 31 类的行为识别准确率稍低,第 33、34 和 38 类的行为识别准确率在 75% 以下,因为这些行为分别为吃饭、摘眼镜、挥手、踢腿、用手指东西、看手表、搓手和敬礼,人体运动部位较少,动作变化幅度较小,行为时空特征较难表征;第 14 类行为是涉及运动部位较多的穿外套行为,人体运动幅度较大,时空特征容易表征,行为的识别准确率稍高,接近 95%。中间 9 类健康相关行为的识别准确率大多在 75% ~ 90% 之间;其中第 44 类头疼的行为,由于动作简单,人体运动部位较少,故识别准确率稍高于 70%。最后 11 类多人交互行为,识别准确率在 75% ~ 95% 之间;第 53 和 54 为拍背和一个人指另一个人的行为,由于人体运动部位较少、运动幅度较小,行为时空特征较难表征,故识别准确率在 80% 以下;第 59 和 60 类是两人面对着走和背对着走的行为,人体运动部位较多,运动幅度较大,故识别准确率较高,在 90% 以上。前 49 类单人行为的平均识别准确率为 82. 43%,后 11 类多人行为的平均识别准确率为 84. 87%,说明本文对于人体运动部位多,运动幅度大的多人行为识别效果更好。

STAR 与其他算法的对比实验结果如表 1 所示,所有方法的识别准确率均不高,因为大多数针对一段

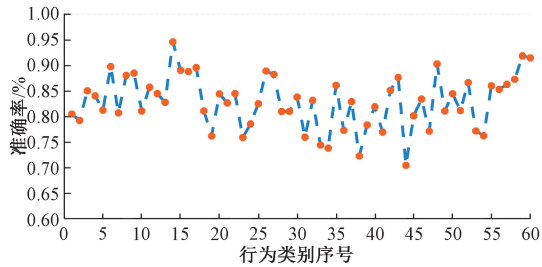


图 8　各行为类别识别准确率

Fig. 8　The recognition accuracy of each action class

骨架序列的行为识别中,输入网络的是 16 帧随机抽取的骨架帧[27,37],而本文仿真中为了模拟人体行为发生时刻和持续时间不确定的情况并构建轻量化网络,输入网络中的是 8 帧关键相邻帧时空特征张量或三维坐标数据,网络的输入信息减少,故识别准确率普遍降低。各方法中,使用 DeepSCNs 识别行为的准确率最低,只有 47. 31%,因为虽然构建 DeepSCNs 网络时通过监督机制保证了网络模型的紧凑,并且提高了模型的表达能力,但是由于 DeepSCNs 对数据时空特性的提取能力较弱,故识别效果不佳。使用 GCN 准确度达到了 66. 54%。因为 GCN 的网络结构有利于人体骨架空间拓扑结构的表示和空间特征的表达,故识别准确率有所上升,但是由于 GCN 缺乏对时序特征的提取故识别准确率仍然不高。LSTM 的识别准确率达到了 75. 67%,比使用 GCN 识别行为的方法精度高出了 9. 13%,证明人体行为识别中对行为时序特征的提取更为关键。使用 CNN 和 LSTM 分别提取人体骨架坐标的空间特征和时间特征的方法,识别准确率达到 81. 89%,但是三维坐标数据对人体行为特征的表征能力有限,网络对行为时空特征的提取不够充分。本文提出的 STAR 识别准确度最高,达到了 84. 57%,证明了本文提出的 STAR 方法良好的性能。

表 1　不同算法识别准确率

Table 1　Recognition accuracy of different algorithms

算法	准确率/%
DeepSCNs	47. 31
GNN	66. 54
LSTM	75. 67
CNN+LSTM	81. 89
STAR	84. 57

2.4 消融实验

为验证本文方法中各模块的有效性，设置消融实验，结果如表 2 所示。

表 2 消融实验结果

Table 2 Ablation results

实验设置	准确率/%
去除相邻帧时空特征构建模块	80.68
去除行为时空特征张量空间注意模块	81.25
去除行为时空特征张量时间注意模块	75.99
使用 FCN 代替 DeepSCNs	81.30
STAR	84.57

去除相邻帧时空特征构建模块后，识别准确率下降了 3.89%。因为其将有利于表征人体骨架时空信息的特征输入网络，而去掉该模块后模型需要通过训练自主探索可以表征人体行为时空信息的特征。去除行为时空特征张量空间注意模块后，识别准确率下降了 2.32%。因为行为时空特征张量空间注意模块使模型更关注重要的空间特征。去除行为时空特征张量时间注意模块后，识别准确率下降了 8.58%。因为人体行为是一个持续的动态过程，需要考虑不同时刻骨架帧之间的关系。同时，与去除行为时空特征张量空间注意模块结果相比识别准确率下降的更多，这也证明了人体行为中的时间特性比空间特性更有价值，表征性更好。行为识别模块中使用 DeepSCNs 识别行为的准确率比使用 FCN 识别行为的准确率高 3.27%。由于监督机制的约束和隐含层与输出层之间的权值连接，DeepSCNs 比 FCN 更加紧致且表达能力更强。

3 应用效果验证

为验证所提 STAR 方法的应用效果，本文开发基于人体骨架的行为识别系统，如图 9 所示。行为识别系统包括硬件和软件两部分。硬件部分主要作用是人体行为数据采集和传输，采用大华高清摄像头，捕获收集完整的行为视频，并通过 RJ-45 接口传输至高性能计算机；软件由骨架数据提取、数据平滑去噪、特征提取、建立模型和行为识别几部分组成，采用骨架提取工具提取骨架坐标数据并平滑去噪后，使用本文所提 STAR 方法提取特征并建立识别模型，最终实现行为识别。

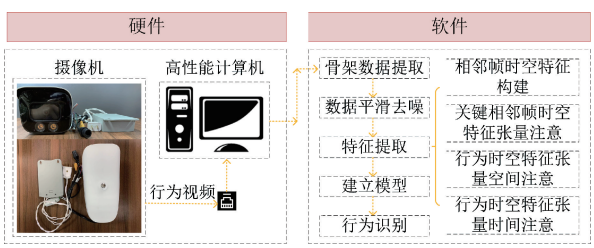


图 9 人体行为识别系统框架图

Fig. 9 Frame diagram of the HAR system

3.1 硬件部分

本文硬件中选用大华的 200 万像素高清摄像头，使用 12 V 恒压电源供电，并通过 RJ-45 接口同计算机通讯。高性能计算机的配置为 CPU i7-11700K 3.6 GHz，RAM 为 16 GB，GPU 为 RTX3060 12 GB。在行为视频数据采集时，研究人员在摄像头摄像覆盖区域进行走路、跑步等行为，摄像机将视频记录并传输至计算机存储。

3.2 软件部分

本部分设计了一款基于人体骨架的行为识别软件。该软件的主要模块包括骨架坐标提取、行为特征提取和行为建模与识别。其中，行为建模与识别最为重要，这里使用本文所提 STAR 算法进行建模与识别。

对于数据采集，本文所使用的数据集来自 9 名志愿者(男女比例为 1∶8)，数据采集过程中他们在摄像头覆盖区域内完成规定的站立、行走、跑步 3 类行为动作，然后摄像头将志愿者的行为视频记录并存储到计算机中。为从行为视频中得到骨架坐标信息，本文使用骨架提取工具 OpenPose[11] 处理采集的行为视频，以便后续进行行为特征提取和构建识别模型。在行为识别模型构建完成后，将其嵌入到软件当中进行实际效果验证。

3.3 效果呈现

本文使用 Qt Creator 和 Python 开发了一款基于人体骨架的行为识别软件，将 OpenPose 骨架提取工具和本文所提 STAR 算法嵌入其中，实现了行为视频、骨架图像、骨架特征、行为结果的实时展示以及存储，实际效果如图 10 所示。行为识别软件运行时通过计算机外接的摄像头实时获取行为视频，之后通过 OpenPose 提取视频中的人体骨架坐标，进而使用 STAR 算法基于

图 10 软件效果图

Fig. 10 Effect drawing of software

坐标数据计算行为特征并识别行为,并且在运行过程中可以选择行为视频、骨架特征等信息的存储位置,以便丰富数据库。

4 结　论

本文针对人体行为的空间复杂性和时间差异性问题提出了一种基于时空张量融合的人体骨架行为自适应识别方法,在相邻帧时空特征构建模块计算相邻帧时空特征张量,关键相邻帧时空特征张量注意模块选取关键相邻帧时空特征张量并组成行为时空特征张量,之后在行为时空特征张量空间注意模块和行为时空特征张量时间注意模块进行特征时空注意,实现行为时空特征融合,最后使用 DeepSCNs 识别人体行为。仿真实验结果表明所提方法在人体行为发生时刻和持续时间不确定情况下,精度高于其他方法,并且设计了基于人体骨架的行为识别系统,验证了所提方法的实际应用效果。但是所使用的数据为无任何遮挡的数据集,故实际情况中根据遮挡数据进行 HAR 是接下来研究工作的重点。

参考文献

[1] WANG J, LIU Z, WU Y, et al. Mining actionlet ensemble for action recognition with depth cameras[C]. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence: IEEE, 2012: 1290-1297.

[2] KONG Y, FU Y. Max-margin action prediction machine[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38: 1844-1858.

[3] CALO R. Robotics and the lessons of cyberlaw[J]. California Law Review, 2015, 103(3): 13-63.

[4] 李瑞峰, 王亮亮, 王珂. 人体动作行为识别研究综述[J]. 模式识别与人工智能, 2014, 27(1): 35-48.
LI R F, WANG L L, WANG K. A survey of human body action recognition[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(1): 35-48.

[5] DANG L M, MIN K, WANG H, et al. Sensor-based and vision-based human activity recognition: A comprehensive survey[J]. Pattern Recognition, 2020, 108: 107561.

[6] 南静, 宁传锋, 建中华, 等. 基于随机配置网络的轻量级人体行为识别模型[J]. 控制与决策, 2023, 38(6): 1541-1550.
NAN J, NING CH F, JIAN ZH H, et al. A lightweight model for human activity recognition using stochastic configuration networks[J]. Control and Decision, 2023, 38(6): 1541-1550.

[7] 周育新, 白宏阳, 李伟, 等. 基于关键帧的轻量化行为识别方法研究[J]. 仪器仪表学报, 2020, 41(7): 196-204.
ZHOU Y X, BAI H Y, LI W, et al. Research on lightweight action recognition method based on key frame[J]. Chinese Journal of Scientific Instrument, 2020, 41(7): 196-204.

[8] 张海超, 张闯. 融合注意力的轻量级行为识别网络研究[J]. 电子测量与仪器学报, 2022, 36(5): 173-179.
ZHANG H CH, ZHANG CH. Research on lightweight action recognition network integrating attention[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(5): 173-179.

[9] WU L F, YANG Z, JIAN M, et al. Global motion estimation with iterative optimization-based independent univariate model for action recognition[J]. Pattern Recognition, 2021: 116.

[10] 孟勃, 刘雪君, 王晓霖. 基于四元数时空卷积神经网络的人体行为识别[J]. 仪器仪表学报, 2017, 38: 2644-2650.
MENG B, LIU X J, WANG X L. Human body action recognition based on quaternion spatial-temporal convolutional neural network[J]. Chinese Journal of Scientific Instrument, 2017, 38: 2644-2650.

[11] 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: 172-186.

[12] SUN Z, KE Q H, RAHMANI H, et al. Human action recognition from various data modalities: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1-20.

[13] 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.
YOU W, WANG X. Study on the edge computing method for skeleton-based human action feature recognition[J]. Chinese Journal of Scientific Instrument, 2020, 41(10): 156-164.

[14] 赵挺, 曹江涛, 姬晓飞. CNN A-BLSTM network 的双人交互行为识别[J]. 电子测量与仪器学报, 2021, 35(11): 100-107.
ZHAO T, CAO J T, JI X F. CNN A-BLSTM network for two-person interaction behavior recognition[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(11): 100-107.

[15] 朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6): 848-857.
ZHU Y, ZHAO J K, WANG Y N, et al. A review of

human action recognition based on deep learning[J]. Acta Automatica Sinica, 2016, 42(6): 848-857.

[16] HUSSEIN M E, TORKI M, GOWAYYED M A, et al. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations [C]. Proceedings of the 2013 International Joint Conference on Artificial Intelligence, Beijing: IJCAI, 2013: 2466-2472.

[17] CIPPITELLI E, GASPARRINI S, GAMBI E, et al. A human activity recognition system using skeleton data from rgbd sensors[J]. Computational Intelligence and Neuroscience, 2016, 3(6): 1-14.

[18] WANG C Y, FLYNN J, WANG Y Z, et al. Recognizing actions in 3d using action-snippets and activated simplices[C]. Proceedings of the 30th AAAI Conference on Artificial Intelligence, Arizona: AAAI Press, 2016: 3604-3610.

[19] KIM H, LEE S, KIM Y, et al. Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system[J]. Expert Systems with Applications, 2016, 45: 131-141.

[20] LI Y, CHU Z J, XIN Y Z. Posture recognition technology based on kinect[J]. IEICE Transactions on Information and Systems, 2020, 103(3): 621-630.

[21] WANG C Y, WANG Y Z, YUILLE A L. Mining 3d key-pose-motifs for action recognition[C]. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas: IEEE, 2016: 2639-2647.

[22] 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 CH Y, LIU K, LI G, et al. Spatio-temporal weighted posture motion features for human skeleton action recognition research [J]. Chinese Journal of Computers, 2020, 43(1): 29-40.

[23] YANG S Q, YANG J T, LI F, et al. Human action recognition based on fusion features[C]. Proceedings of the 2022 International Conference on Cyber Security Intelligence and Analytics, Haikou: CSIA, 2020, 928: 569-579.

[24] LIANG X Y, ZHANG H B, ZHANG Y X, et al. JTCR: Joint trajectory character recognition for human action recognition[C]. Proceedings of the 2019 IEEE Eurasia Conference on IOT, Communication and Engineering, Yunlin: IEEE, 2019: 350-353.

[25] DAI C, WEI Y J, XU Z J, et al. An investigation of gcn-based human action recognition using skeletal features [C]. Proceedings of the 27th International Conference on Automation and Computing, Bristol: CACSUK, 2022: 1-10.

[26] ZHANG H B, ZHANG Y X, ZHONG B E, et al. A comprehensive survey of vision-based human action recognition methods [J]. Sensors (Basel), 2019, 19(5): 1005.

[27] ANGELINI F, FU Z, LONG Y, et al. 2D pose-based real-time human action recognition with occlusion-handling[J]. IEEE Transactions on Multimedia, 2019, 99:1,DOI:10.1109/TMM.2019.2944745.

[28] ZHU X G, ZHU Y, WANG H Y, et al. Skeleton sequence and RGB frame based multi-modality feature fusion network for action recognition [J]. ACM Transactions on Multimedia Computing, 2022, 18(3): 1-24.

[29] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]. Proceedings of the 2017 International Conference on Learning Representations, Toulon: ICLR, 2017.

[30] FU B, FU S L, WANG L Y, et al. Deep residual split directed graph convolutional neural networks for action recognition[J]. IEEE Multi Media, 2020, 27(4): 9-17.

[31] 薛盼盼, 刘云, 李辉, 等. 基于时域扩张残差网络和双分支结构的人体行为识别[J]. 控制与决策, 2022, 37(11): 2993-3002.
XUE P P, LIU Y, LI H, et al. Human behavior recognition based on time domain extended residual network and dual branching structure[J]. Control and Decision, 2022, 37(11): 2993-3002.

[32] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans: AAAI, 2018:1-10.

[33] BAI S J, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling [J]. ArXiv Preprint, 2018, ArXiv:1803.01271.

[34] LI M S, CHEN S H, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition [C]. Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach: IEEE, 2019.

[35] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach: NIPS, 2017: 6000-6010.

[36] MAZZIA V, ANGARANO S, SALVETTI F, et al. Action transformer: A self-attention model for short-time pose-based human action recognition [J]. Pattern Recognition, 2021: 124, DOI:10.1016/j.patcog.2021.108487.

[37] PLIZZARI C, CANNICI M, MATTEUCCI M. Skeleton-based action recognition via spatial and temporal transformer networks [J]. Computer Vision and Image Understanding, 2020: 208-209.

[38] ZHANG Y X, LI B H, FANG H, et al. Current advances on deep learning-based human action recognition from videos: A survey [C]. Proceedings of the 20th IEEE International Conference on Machine Learning and Applications, Pasadena: AMLA, 2021: 304-311.

[39] WANG P C, LI W Q, 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.

[40] WANG D H, LI M. Stochastic configuration networks: Fundamentals and algorithms [J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3466-3479.

[41] DAI W, LI D P, ZHOU P, et al. Stochastic configuration networks with block increments for data modeling in process industries [J]. Information Sciences, 2019, 484: 367-386.

[42] DAI W, ZHOU X Y, LI D P, et al. Hybrid parallel stochastic configuration networks for industrial data analytics [J]. IEEE Transactions on Industrial Informatics, 2022, 18(4): 2331-2341.

[43] WANG Q J, HONG Q Q, WU S, et al. Multi-target stochastic configuration network and applications [J]. IEEE Transactions on Artificial Intelligence, 2022, 18: 2331-2341.

[44] WANG D H, LI M. Deep stochastic configuration networks with universal approximation property [C]. 2018 International Joint Conference on Neural Networks. Rio de Janeiro: IEEE, 2018: 1-8.

[45] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis [J]. IEEE Computer Vision and Pattern Recognition, 2016: 1010-1019.

作者简介

建中华,2019 年于齐齐哈尔大学获得学士学位,2023 年于中国矿业大学获得硕士学位,现于中国矿业大学攻读博士学位,主要研究方向为人体行为识别,机器学习,数据分析。

E-mail: 18845215747@163.com

Jian Zhonghua received his B. Sc. degree from Qiqihar University in 2019, and received his M. Sc. degree from China University of Mining and Technology in 2023. He is currently a Ph. D. candidate at China University of Mining and Technology. His main research interests include human action recognition, machine learning and data analysis.

代伟(通信作者),2015 年毕业于东北大学流程工业综合自动化国家重点实验室,获得博士学位。现为中国矿业大学人工智能研究院、信息与控制工程学院教授,博导,主要研究方向为工业数据分析与建模、工业运行优化控制。

E-mail: weidai@cumt.edu.cn

Dai Wei (Corresponding author) received his Ph. D. degree from State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University in 2015. He is currently a professor and a Ph. D. advisor in the Artificial Intelligence Research Institute and School of Information and Control Engineering, China University of Mining and Technology. His main research interests are data analysis and modeling, industrial operation optimization control.