Journal of System Simulation

Volume 35 | Issue 5 Article 17

5-22-2023

Action Recognition Method Based on Projection Subspace Views under Single Viewing Angle

Benyue Su
1.The Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing Normal University, Anqing 246133, China;2.School of Mathematics and Computer, Tongling University, Tongling 244061, China;, subenyue@sohu.com

Manzhen Sun
3.School of Computer and Information, Anqing Normal University, Anqing 246133, China;

Qing Ma
4.School of Mathematics and Physics, Anqing Normal University, Anqing 246133, China;

Min Sheng
4.School of Mathematics and Physics, Anqing Normal University, Anqing 246133, China;

Action Recognition Method Based on Projection Subspace Views under Single Viewing Angle

Abstract

Abstract: In view of the self-occlusion problem of joint action tracking by a depth camera under a single viewing angle, *a new human action recognition method based on projection subspace views is proposed. Without adding data acquisition equipment, the method projects the three-dimensional*(*3D*) *action sequences obtained under a single viewing angle into multiple two-dimensional subspacesand then seeks the maximum distance between classes in the two-dimensional subspaces, so as to increase the distance between 3D actions based on the fusion of multiple subspace views as much as possible.* The recognition rate in the self-built AQNU dataset is 99.69%, which is 1.22% higher than the benchmark method. The recognition rate in the public NTU-RGB+D dataset subset is 80.23%, which is 1.98% higher than the benchmark method. The experimental results show that the method proposed in this paper can alleviate the self-occlusion problem of datasets of single viewing angles to a certain extent, effectively improve the recognition rate and computational efficiency, and achieve the recognition effect equivalent to that of datasets of multiple viewing angles.

Keywords

action recognition, single view, projection subspace, graph convolutional network

Recommended Citation

Benyue Su, Manzhen Sun, Qing Ma, Min Sheng. Action Recognition Method Based on Projection Subspace Views under Single Viewing Angle[J]. Journal of System Simulation, 2023, 35(5): 1098-1108.

第 35 卷第 5 期 系统仿真学报© Vol. 35 No. 5
2023 年 5 月 Journal of System Simulation May 2023

单视角下基于投影子空间视图的动作识别方法

苏本跃[1,2]，孙满贞[3]，马庆[4]，盛敏[4]

(1. 安徽省智能感知与计算重点实验室，安庆师范大学，安徽 安庆 246133；2. 铜陵学院 数学与计算机学院，安徽 铜陵 244061；
3. 安庆师范大学 计算机与信息学院，安徽 安庆 246133；4. 安庆师范大学 数理学院，安徽 安庆 246133)

摘要：针对单视角下深度相机跟踪关节点运动存在的自遮挡问题，*提出一种基于投影子空间视图的人体动作识别方法。在不增加数据采集设备的情况下，通过子空间投影，将单视角下获得的三维动作序列投影到多个二维子空间中，在二维投影空间寻求最大类间距离，以尽可能增加基于多个子空间视图融合后的3D动作类间距离。*在自建AQNU数据集的识别率为99.69%，较基准方法提升1.22%。在公共NTU-RGB+D数据集子集的识别率为80.23%，较基准方法提升1.98%。实验结果表明：本文方法可在一定程度上减少单视角数据集的自遮挡问题，提高识别率和计算效率，可达到与多视角数据集相当的识别效果。

关键词：动作识别；单视图；投影子空间；图卷积网络

中图分类号：TP391.41；TP391.9　　文献标志码：A　　文章编号：1004-731X(2023)05-1098-11

DOI: 10.16182/j.issn1004731x.joss.22-0087

引用格式: 苏本跃, 孙满贞, 马庆, 等. 单视角下基于投影子空间视图的动作识别方法[J]. 系统仿真学报, 2023, 35(5): 1098-1108.

Reference format: Su Benyue, Sun Manzhen, Ma Qing, et al. Action Recognition Method Based on Projection Subspace Views under Single Viewing Angle[J]. Journal of System Simulation, 2023, 35(5): 1098-1108.

Action Recognition Method Based on Projection Subspace Views under Single Viewing Angle

Su Benyue[1,2], *Sun Manzhen*[3], *Ma Qing*[4], *Sheng Min*[4]

(1. The Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing Normal University, Anqing 246133, China; 2. School of Mathematics and Computer, Tongling University, Tongling 244061, China; 3. School of Computer and Information, Anqing Normal University, Anqing 246133, China; 4. School of Mathematics and Physics, Anqing Normal University, Anqing 246133, China)

Abstract: In view of the self-occlusion problem of joint action tracking by a depth camera under a single viewing angle, *a new human action recognition method based on projection subspace views is proposed. Without adding data acquisition equipment*, *the method projects the three-dimensional*(*3D*) *action sequences obtained under a single viewing angle into multiple two-dimensional subspaces and then seeks the maximum distance between classes in the two-dimensional subspaces*, *so as to increase the distance between 3D actions based on the fusion of multiple subspace views as much as possible.* The recognition rate in the self-built AQNU dataset is 99.69%, which is 1.22% higher than the benchmark method. The recognition rate in the public NTU-RGB+D dataset subset is 80.23%, which is 1.98% higher than the benchmark method. The experimental results show that the method proposed in this paper can alleviate the self-occlusion problem of datasets of single viewing angles to a certain extent, effectively improve the recognition rate and computational efficiency, and achieve the recognition effect equivalent to that of datasets of multiple viewing angles.

Keywords: action recognition; single view; projection subspace; graph convolutional network

收稿日期：2022-01-27　　修回日期：2022-04-27

基金项目：安徽省自然科学基金(2108085QF269)；高校领军人才团队项目(皖教秘人[2019]16号)

第一作者：苏本跃(1971-)，男，教授，博士，研究方向为图形图像处理、机器学习与模式识别等。E-mail：subenyue@sohu.com

0 引言

近年来基于骨骼数据的人体动作识别是计算机视觉领域的研究热点之一，广泛应用于人机交互、智能安防和无人驾驶等领域[1-3]。骨骼数据是人体关节点的3D坐标序列，受光照、背景和视角变换的影响较小，可视为人体运动的高级特征[4]。且早在1973年，约翰逊[5]的生物学观察表明，在没有外观信息的条件下，人体动作可以用几个关节点坐标位置的变化表示。在实际应用中，单视角下的人体行为监控广泛应用于智能看护和公共安全防护领域。

现阶段基于深度学习的人体动作识别常用的深度网络有循环神经网络(recurrent neural network, RNN)、卷积神经网络(convolutional neural network, CNN)、图卷积网络(graph convolutional network, GCN)。RNN、CNN和GCN分别将骨骼数据表征为长向量、伪图像和拓扑图。鉴于人体骨骼数据可视为由节点(关节)和边(骨骼)构成的图数据，而GCN能够直接处理由点和边组成的图数据，同时学习节点自身的特征、相邻节点以及非相邻节点之间的关系用于分类任务，因此在人体动作识别问题中，GCN相较于RNN与CNN，更能利用人体骨骼的拓扑结构信息进行特征学习与动作识别。文献[6]提出的时空图卷积网络(spatial temporal graph convolutional network, ST-GCN)首次将GCN引入基于骨骼数据的人体动作识别领域，此后，基于ST-GCN模型改进的工作不断出现[7-11]。鉴于ST-GCN模型可以直接对动作的原始3D骨骼数据进行学习，无需将数据转化为长向量或伪图像；且它能够从人体动态骨骼数据自动学习动作在空间和时间方面变化的信息，实现了信息在时空维度的集成；更重要的是，ST-GCN模型在多个公开数据集的识别结果较RNN和CNN均有质的提升。因此，本文以ST-GCN模型为基准，提出融合动作三视图的时空图卷积网络(three views fusion spatial temporal graph convolutional networks, TVF-STGCN)。

动作观察视角是人体动作识别中另一个值得思考的问题，也是人体动作识别领域现阶段存在的挑战之一[12]。研究者基于骨骼数据动作识别的研究逐渐从单视角[13-16]过渡到多视角[17-19]。单视角是运用一个相机采集的数据进行动作识别，其优势是数据采集便利、处理简单和计算量小、但其缺点是从固定的一个视角观察动作存在关节点自遮挡现象[20]，这不利于准确采集动作数据，从而影响动作识别效果。多视角是运用多个相机(>1)同时观察动作，其优点是不同方位的相机获得的动作序列信息互补，且能在一定程度上解决单视角观察动作存在的自遮挡问题。但其缺点是数据采集的成本高、对数据采集的环境要求高、数据量大且计算消耗多。为了利用单视角数据采集成本低的优势，同时达到多视角动作数据集的识别效果，本文提出单视角下融合动作多视图的识别方法。

针对单视角深度监控相机对人体进行动作数据采集时存在的关节点自遮挡问题，本文提出单视角下基于投影子空间视图的动作识别方法。具体地，选择3D动作投影空间中的3个互相垂直的坐标面作为动作的投影子空间，这3个坐标面是3D动作三视图的投影平面。在机械工程制造中，根据三维刚体的三视图可以完成3D物体的重建[21]。借鉴此思想用于3D骨骼数据动作识别，根据3D动作的三视图完成3D动作的重构，并进一步识别出动作标签。

综合考虑以上因素，本文提出单视角下融合动作三视图的时空图卷积网络。具体地，首先根据Kinect相机坐标系定义3D动作的主视图、左视图和俯视图，得到3个视图的2D动作序列。然后用ST-GCN对不同视图的动作序列进行特征学习，分别获取单个视图的动作识别标签。最后运用投票策略对3个视图的动作识别标签投票，得到模型的最终识别结果。

本文的主要贡献如下：①提出一种单视角下

DOI: 10.16182/j.issn1004731x.joss.22-0087

融合动作三视图时空图卷积网络(TVF-STGCN)模型，实验结果表明，该模型有效提高了动作识别精度。②基于三视图的思想，将三维空间动作识别问题转化为3个二维子空间的动作识别问题，同时降低了数据维度和模型的计算复杂度。③运用投票策略融合3个视图的动作识别标签，能综合利用动作在3个视图的互补信息，从而对动作进行精准分类。

1 模型和方法

1.1 预备知识

骨骼序列数据是一系列帧数据，在一帧中构建时空无向图$G(V, E)$，其中V表示关节点集，E表示骨骼边集。$V=\{V_{ti}|t=1, 2, \cdots, T, i=1, 2, \cdots, N\}$包含动作序列中所有关节点坐标数据，$T$表示总帧数，$N$表示关节点总数，$V_{ti}$表示第$i$个关节点在第$t$帧时的坐标。骨骼边集$E$包含2个子集，第1个子集表示一帧之内相邻关节点连接形成的边，称为帧内边(空间边)；第2个子集表示相邻帧同一个关节点连接形成的边，称为帧间边(时间边)。帧内边表示动作空间结构信息的变化，帧间边表示动作时间动态信息的变化。以挥手动作为例，其动作序列时空图如图1所示。

1.2 单视角下融合3D动作三视图的人体动作识别

单视角下观察动作时由于视野范围受限存在关节点自遮挡问题，本文提出融合动作三视图的方法从3个方向对同一3D动作进行投影，综合利用3个视图的2D动作信息，能够有效解决这类问题。具体地，动作的主视图从动作的正前方水平向后投影，反映动作正前方的运动状态；动作的左视图从动作的左方向右投影，反映动作左方的运动状态；动作的俯视图从动作的正上方向下投影，反映动作正上方的运动状态；单视角下融合三视图方法的动作识别总体流程如图2所示。

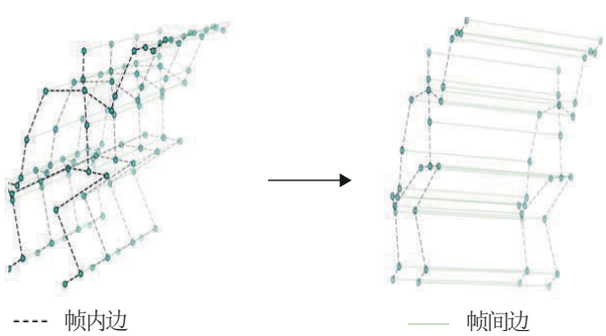


图1 以挥手动作为例展现动作序列的时空图
Fig. 1 Spatiotemporal diagram of action sequences with hand waving as an example

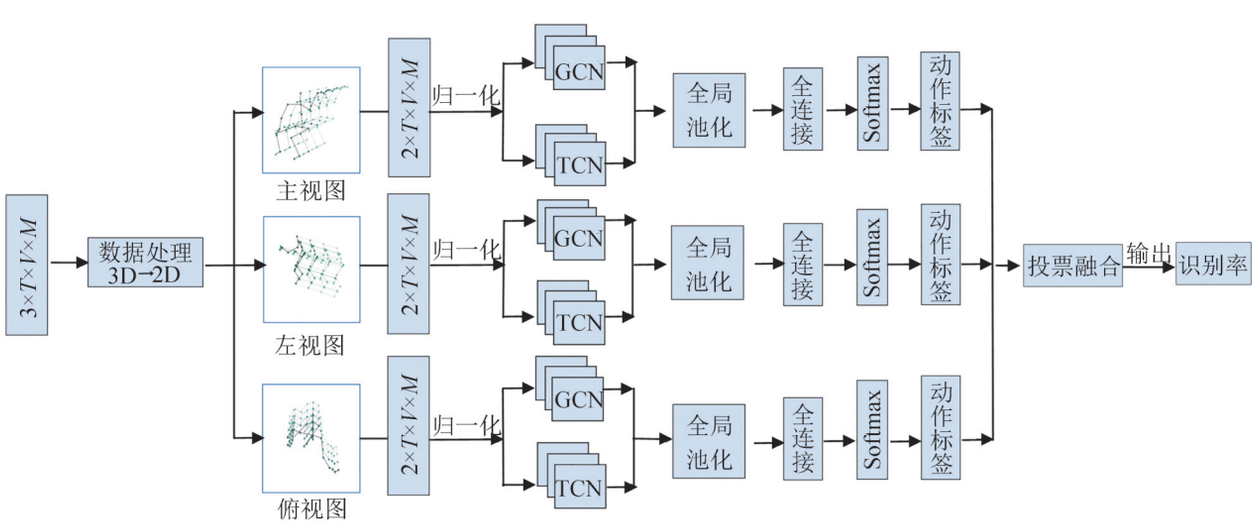


图2 融合动作三视图时空图卷积网络框架
Fig. 2 Spatiotemporal graph convolutional network framework fusing three views of actions

从图2可知，融合动作三视图时空图卷积网络的工作主要包括3个基本模块。

(1) 数据处理模块。原始3D骨骼序列数据可以表示为3×T×V×M，3表示关节点坐标维数，T表示动作样本帧数，V表示关节点数，M表示动作样本中的人数，经过投影处理后得到3个2D子空间的动作序列，数据格式为2×T×V×M，映射的具体过程如图3所示。

(2) 单视图动作识别模块。首先将3D动作的主视图、左视图和俯视图动作序列作归一化处理，然后分别输入ST-GCN并行计算得到3个视图的动作识别标签。

(3) 融合策略模块。运用投票策略对3个视图的动作标签进行融合，输出模型最终识别结果。

本文TVF-STGCN算法流程如下：

输入：

将3D动作通过投影得到的3个二维平面的动作序列分别输入ST-GCN。数据格式为(N ×M, C, V, T)

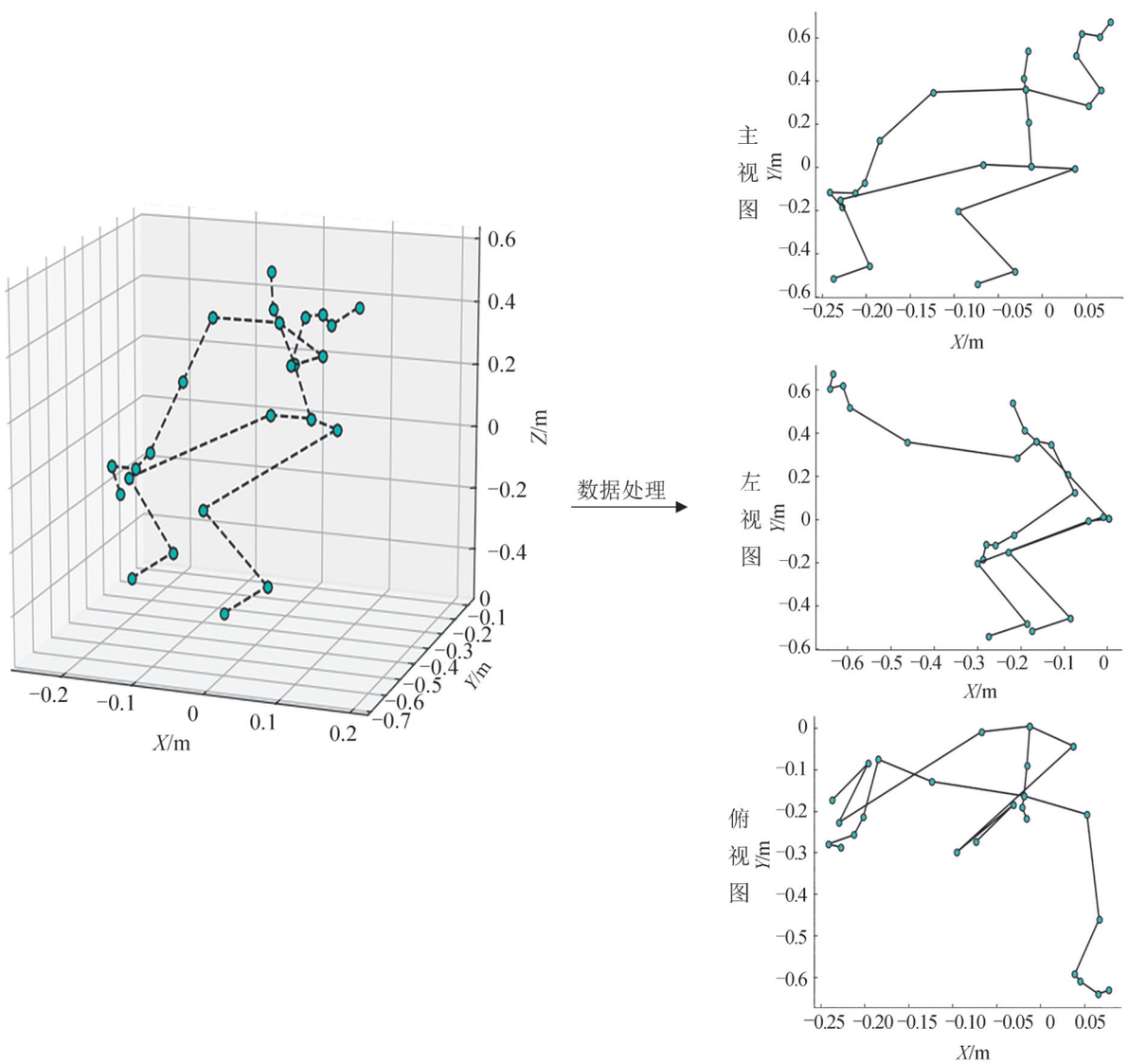


图3　以挥手动作为例的数据处理示意图

Fig. 3　Schematic diagram of data processing with hand waving as an example

输出：

融合之后的识别率

主体框架：

step 1：对输入矩阵在时空维度进行归一化；

step 2：通过ST-GCN单元，交替使用GCN和TCN，对时间和空间维度进行变换；

step 3：给定一个Kernel size的值K,在位置x的单个通道的输出值为$f_{out}(x)$；

step 4：重新定义采样函数P，定义节点V_{ti}的邻域集$B(V_{ti})=\{V_{tj}|d(V_{tj},V_{ti})\leqslant D\}$，设定$D$=1；

step 5：重新定义权重函数W，采用空间分布分区策略，将一个关节点V_{ti}的邻域集$B(V_{ti})$划分为固定数量的3个子集来简化该过程；

step 6：考虑动作序列在时间上的卷积，关节点V_{ti}的时间邻域记为$l_{st}(V_{qj})$，卷积从空域转到时空域；

step 7：使用全局池化层和全连接层对卷积得到的特征进行分类，输出单个视图的动作标签；

step 8：使用投票策略对三视图输出的动作标签进行投票，得到模型最终识别结果。

1.3 三视图中具体一个视图的动作识别

以主视图骨骼序列数据为例，原始的3D骨骼序列沿着XOY平面投影得到主视图的2D动作序列。归一化处理后输入一个包含9个时空卷积模块的ST-GCN，得到主视图的动作识别标签。具体的识别流程如图4所示。

由图4可知，主视图的动作序列数据在归一化处理后，通过9个时空图卷积模块提取动作序列的时空特征。其中每个模块均包含注意力块、图卷积

块和时间卷积块。卷积后得到256维的特征向量，经过SoftMax层输出动作标签。

空间卷积中，单通道节点V_{ti}的输出值为

$$f_{\text{out}}(V_{ti})=\sum_{V_{tj}\in B(V_{ti})}\frac{1}{Z_{tj}(V_{tj})}f_{\text{in}}(V_{tj})W(l_{ti}(V_{tj})) \quad (1)$$

式中：f_{in}表示输入特征图；$B(V_{ti})$为节点V_{ti}的邻居集，$B(V_{ti})=\{V_{tj}|d(V_{tj},V_{ti})\leqslant D\}$，$D=1$表示节点$V_{ti}$的一阶邻居，将$B(V_{ti})$划分为$K$个子集；$l_{ti}(V_{tj})$为将节点$V_{tj}$映射到节点$V_{ti}$邻居子集的映射，$l_{ti}$: $B(V_{ti})\rightarrow\{0,1,\cdots,K-1\}$；$Z_{tj}(V_{tj})$为归一化项，用于平衡不同子集的贡献；$W$为权重函数。

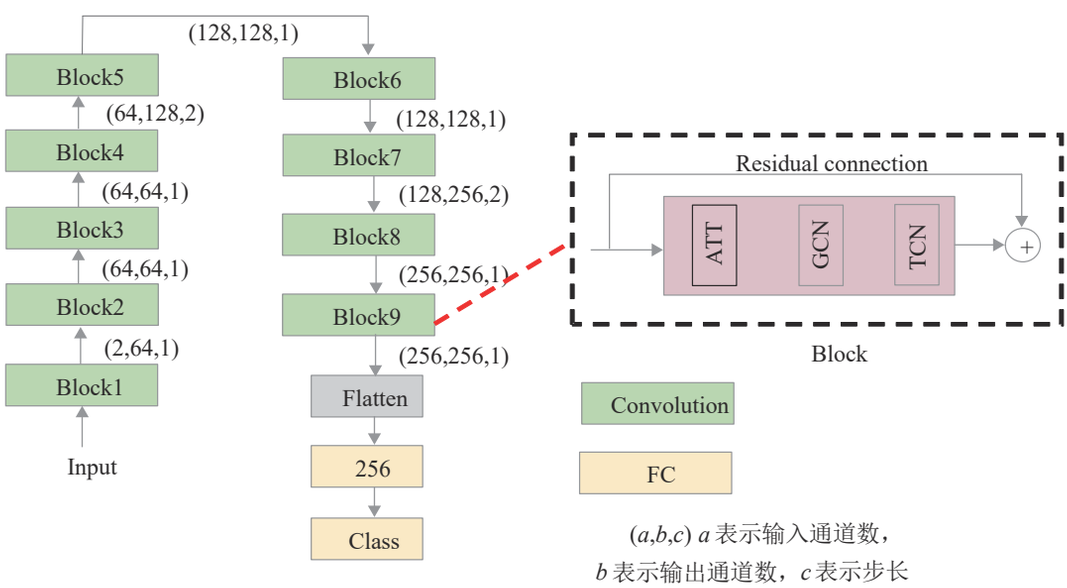


图4 主视图动作序列结合ST-GCN模型的识别流程

Fig. 4 Recognition process of front view action sequence by ST-GCN model

在每两个时空模块，输出值为

$$f_{\text{out}}=\sum_j \Lambda_j^{-\frac{1}{2}}\boldsymbol{A}_j\Lambda_j^{-\frac{1}{2}}f_{\text{in}}\boldsymbol{W}_j \tag{2}$$

式中：f_{in}为输入特征图；$\boldsymbol{A}$为邻接矩阵，I为节点自身的连接，$\boldsymbol{A}+I=\sum_j \boldsymbol{A}_j$；$\Lambda_j^{ii}=\sum_K(\boldsymbol{A}_j^{ik}+\partial)$，设置$\partial=0.001$避免$\boldsymbol{A}_j$中的行出现0；$\boldsymbol{W}_j$为$\boldsymbol{A}_j$的权重矩阵。然后通过全局池化层和全连接层，每个动作序列得到一个256维的特征向量，将其输入SoftMax进行分类。

1.4 投票融合策略

根据1.2节单视角下融合3D动作三视图并行处理的整体识别框架可知，主、左、俯每条支路经过ST-GCN学习分别输出一个动作标签，本文采用投票融合策略获得最终动作识别标签。具体地，对3个通道输出的动作标签进行直方图统计，从而得到最终动作标签。其中对于完全均衡(无法得到最终结果)的直方图，以主视图支路的识别结果为准。

2 数据集和实验

2.1 数据集分析

为了更加贴近深度相机在家庭和公共场合的实际应用场景，本文自建了AQNU数据集。本文模拟这些场合相机安装的位置，把相机固定在实验者的斜上方(45°)，使得相机采集的人体动作数据是非正对着相机执行的，目的是不约束正面采集模式。同时，为了验证本文方法识别性能的鲁棒性，本文借用了公共数据集NTU-RGB+D中的1号相机采集的样本进行试验，因为1号相机同样从人体45°方向观察人体动作，同时相机的高度不固定 (0.5~2.7 m)。

(1) 自建AQNU数据集

在AQNU数据集中，本文共招募了49名健康实验者(女21名，男28名)，年龄18~29岁，身高158~183 cm，体重45~85 kg。该数据集包含20个室内日常动作，每个实验者执行每个动作5次，共4 900个样本。相机在实验者左上方45°的位置，其高度为1.95 m，与实验者的水平距离为3.2 m。具体采集环境如图5所示，动作名称见表1。

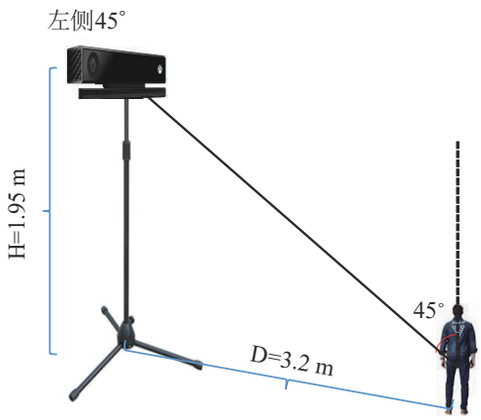


图5 自建AQNU数据集采集环境
Fig. 5 Acquisition environment of self-built AQNU dataset

表1 自建AQNU数据集中包含的动作名称
Table 1 Action names in self-built AQNU dataset

编号	动作	编号	动作
1	喝水	11	欢呼
2	坐下	12	双手交叉置于胸前
3	抱头转头	13	顺时针转圈走
4	扫地	14	摆手臂
5	鼓掌	15	双手捶打
6	抬右腿	16	体侧运动
7	双手戴眼镜	17	扭手
8	投篮球	18	顺时针画圈
9	原地踏步走	19	踢毽子
10	慢跑	20	扩胸运动

(2) NTU-RGB+D子集

NTU-RGB+D数据集[17]采用3个分别放置在实验者45°、0°、−45°方向的Kinect 2.0深度相机同时进行数据采集，是个多视角动作数据集。考虑到NTU-RGB+D的1号相机基于人体45°方向、且相机高度高于实验者身高的设置下采集人体动作数据。本文称NTU-RGB+D的1号相机采集的样本为NTU-RGB+D子集，它是个单视角动作数据集。为进一步验证本文方法的有效性，在NTU-RGB+D子集进行实验。1号相机从45°方向采集实

验者的动作样本，相机高度变化为0.5~2.7 m。

2.2 实验策略及环境

在AQNU数据集中，每个实验者的前4遍动作作为训练集，第5遍动作作为测试集。在NTU-RGB+D子集中，每个实验者每个设置号下的第1遍动作作为训练集，第2遍动作作为测试集。

3个视图子网络都是从0开始训练的。Batch_size设置为256，损失函数采用交叉熵，优化算法为Adam，epoch设置为80，学习率为0.000 1，其余参数均为默认值。使用Python 3.7编程，基于Pytorch 0.4框架在一台配有4块RTX 5000GPU、256 G内存、2.30 GHz处理器以及Ubuntu18.04.3 LTS操作系统的服务器执行实验。

2.3 实验结果及分析

2.3.1 在AQNU数据集的实验结果

本文方法在AQNU数据集的实验结果如表2所示。

由表2可知本文方法在AQNU数据集的俯视图、左视图和主视图的识别率分别为97.99%，98.09%，98.67%，融合三视图的识别率达到99.69%。主视图的识别率高于左视图和俯视图的识别率，原因是从正面观察人体动作时，骨骼关节点以及骨骼之间的自遮挡较少、关节点的坐标受噪声干扰较小、较为准确、故识别效果较好。而从左视图视角和俯视图视角观察动作时，人体关节点之间的自遮挡和重叠情况严重，动作标签辨别的难度增加，故识别效果较主视图略差。融合三视图的识别率高于任何一个单视图的识别率，原因是综合多个视图的动作信息可以减少单视角视野盲区带来的影响，有利于准确识别动作。特别地，在AQNU数据集，俯视图动作序列的识别效果也较好，达到97.99%。其原因是在采集数据时，相机的摆放高度是1.95 m，超过所有实验者的身高，使得相机较为准确地采集到动作俯视图的动作序列，即俯视图在识别动作时也能发挥作用。最后，对比直接运用3D动作序列结合ST-GCN的识别率，本文方法的识别率提高1.22%；模型训练时间减少18 min，测试时间减少14 s，模型训练速度提升13.6%，模型测试速度提升37.8%，即本文方法在计算效率方面有一定的提升。算法计算效率提升的具体原因是：本文算法主要是对传统ST-GCN的重构，将已有的3D骨骼数据结合ST-GCN进行动作识别的串行计算，变为融合动作三视图三条支路分别结合ST-GCN进行动作识别的并行计算。从单条支路看，每条支路数据维度降低使得支路的计算效率得以提升。从模型整体看，动作三视图三条支路并行计算，然后在支路尾端融合不会造成额外的时间消耗。综上表明，融合3D动作三视图用于动作识别，在一定程度上减少了单视角下骨骼关节点自遮挡，提高了计算效率。

2.3.2 在NTU-RGB+D子集的实验结果

为了进一步验证融合3D动作三视图思想用于动作识别的有效性，本文借用NTU-RGB+D数据集中1号相机采集的样本作为NTU-RGB+D数据集的子集，1号相机始终从人体45°方向采集动作，使得3D动作的主视图、左视图和俯视图对于识别人体动作都有效。在NTU-RGB+D子集的实验结果见表3。

表2 本文方法在AQNU数据集的实验结果

Table 2 Experimental results of proposed method in AQNU dataset

实验方法	单个视图的识别率/%	模型整体识别率/%	模型训练时间/min	模型测试时间/s
ST-GCN[6]	—	98.47(3D坐标)	132	37
本文方法	俯视图：97.99 左视图：98.09 主视图：98.67	99.69(融合三视图)	114	23

表3 本文方法在NTU-RGB+D子集的实验结果
Table 3 Experimental results of proposed method in NTU-RGB+D subset

实验方法	单个视图的识别率/%	模型整体识别率/%	模型训练时间/min	模型测试时间/s
ST-GCN[6]	—	78.25 (3D坐标)	193	119
本文方法	俯视图:53.63 左视图:65.80 主视图:78.10	80.23 (融合三视图)	165	97

由表3可知，本文方法在NTU-RGB+D子集的俯视图、左视图、主视图的识别率分别为53.63%、65.80%和78.10%，运用融合策略后的识别率为80.23%。主视图的识别率高于左视图的识别率，左视图识别率的高于俯视图的识别率，融合三视图后的识别率高于任何一个单视图的识别率。本文方法的识别率比直接运用3D骨骼序列结合ST-GCN的识别率提升1.98%，模型训练时间减少28 min，模型测试时间减少22 s，即模型的训练速度和测试速度分别提升14.5%和18.5%。在公共数据集NTU-RGB+D子集的实验结果进一步验证了本文方法的有效性。

本文方法在AQNU数据集的识别率比NTU-RGB+D子集的识别率高19.46%。主要原因是AQNU数据集包含的动作种类少，且对动作执行的规范性要求高，使得同种动作类内差异性小。同时自建AQNU数据集中，实验者的年龄、身高和体重变化范围都较小，有助于准确识别动作标签。

2.3.3 与最新方法的比较

本文方法在NTU-RGB+D数据集的识别结果与其他方法的实验结果对比见表4，从中可以看出本文方法的有效性。

由表4可知，本文方法在NTU-RGB+D子集(单视角)得到的实验结果与表中文献方法在NTU-RGB+D数据集(多视角)得到的实验结果相比总体略低。这是因为NTU-RGB+D数据集本身是多视角数据集，多视角下获得的动作信息互补；多视角观察动作能减少关节点自遮挡现象带来的影响；NTU-RGB+D数据集的样本量是NTU-RGB+D子集数据量的三倍。但是本文方法基于单视角数据集的识别率达到了与多视角数据集相当的识别效果。

2.4 消融实验–用户独立下的实验结果及分析

用户独立策略下，本文使用一半实验者的全部样本作为训练集，另一半实验者的全部样本作为测试集。

表4 本文方法与其他方法在NTU-RGB+D数据集的实验结果对比
Table 4 Comparison of experimental results between proposed method and other methods in NTU-RGB+D dataset

实验方法	识别率/%	数据集	视角
Geometric Features [22]	82.40	NTU-RGB+D	M
TCN [23]	83.10	NTU-RGB+D	M
CNN+MTLN [24]	84.80	NTU-RGB+D	M
ST-LSTM+Trust-Gate [25]	77.70	NTU-RGB+D	M
Deep STGCK [26]	86.30	NTU-RGB+D	M
GCA-LSTM [27]	84.00	NTU-RGB+D	M
TSSI+GLAN+SSAN [28]	89.10	NTU-RGB+D	M
PA-GCN [29]	82.70	NTU-RGB+D	M
ST-GCN [6]	78.25	NTU-RGB+D Subset	S
本文方法	80.23	NTU-RGB+D Subset	S

注：M表示多视角(multi camera)，S表示单视角(single camera)。

DOI: 10.16182/j.issn1004731x.joss.22-0087

2.4.1 在AQNU数据集的实验结果

用户独立策略下，本文方法在AQNU数据集的实验结果见表5。

由表5可知，用户独立策略下，本文方法在AQNU数据集的俯视图、左视图、主视图的实验结果分别为93.12%、96.75%、97.25%，融合三视图的识别率为98.75%。融合三视图的识别率较主视图识别率提高1.50%，且融合三视图的识别率较直接运用3D动作序列结合ST-GCN的识别率提高3.17%，模型训练时间减少15 min，模型测试时间减少19 s，即模型的训练速度和测试速度分别提升12.9%和33.9%，进一步表明了本文方法的有效性。

2.4.2 在NTU-RGB+D子集的实验结果

用户独立策略下，本文方法在NTU-RGB+D子集的实验结果见表6。

由表6可以看出，用户独立策略下，本文方法在NTU-RGB+D子集俯视图、左视图、主视图的实验率分别为52.62%、64.96%、76.42%，融合三视图的识别率为77.57%。融合三视图的识别率较主视图提高1.15%，且融合三视图的识别率较直接运用3D动作序列结合ST-GCN的识别率提高2.93%，模型训练时间减少27 min，模型测试时间减少24 s，即模型的训练速度和测试速度分别提升11.2%和25.2%。这表明了本文方法的有效性。

2.3节与2.4节的实验结果对比如图6所示。

表5 用户独立策略下—本文方法在AQNU数据集的识别结果

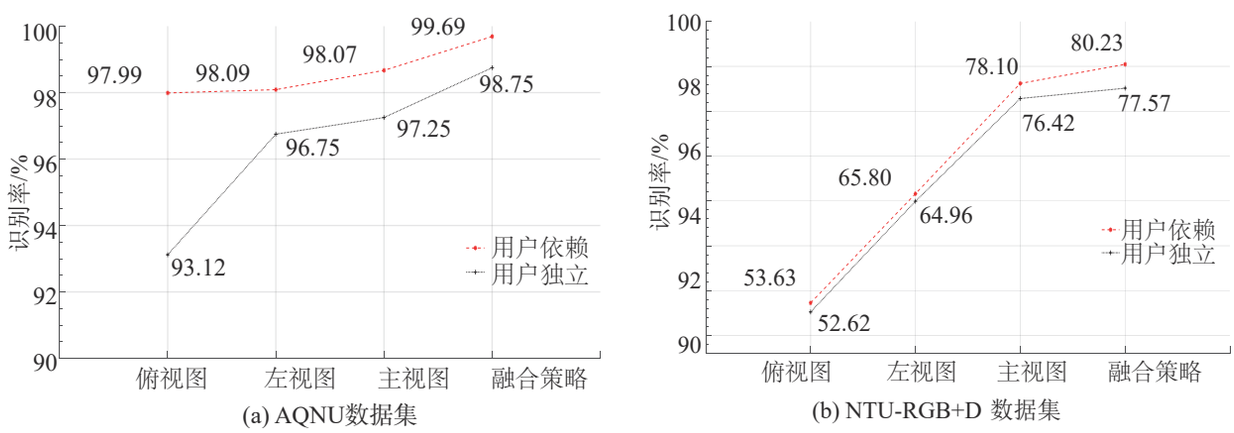
Table 5 Recognition results of proposed method in AQNU dataset under user-independent strategy

实验方法	单个视图的识别率/%	模型整体识别率/%	模型训练时间/min	模型测试时间/s
ST-GCN[6]	—	95.58 (3D坐标)	116	56
本文方法	俯视图:93.12 左视图:96.75 主视图:97.25	98.75 (融合三视图)	101	37

表6 用户独立策略下—本文方法在NTU-RGB+D子集的识别结果

Table 6 Recognition results of proposed method in NTU-RGB+D subset under user-independent strategy

实验方法	单个视图的识别率/%	模型整体识别率/%	模型训练时间/min	模型测试时间/s
ST-GCN[6]	—	74.64 (3D坐标)	241	95
本文方法	俯视图:52.62 左视图:64.96 主视图:76.42	77.57 (融合三视图)	214	71



(a) AQNU数据集 (b) NTU-RGB+D 数据集

图6 两种实验策略下数据集的实验结果对比

Fig. 6 Comparison of experimental results in datasets under two experimental strategies

观察图6，通过对比2.3节和2.4节的实验结果，可以发现在两个数据集，2.3节的实验结果不论是在单个视图还是融合三视图之后，其识别率均高于2.4节的识别率。造成这种现象的根本原因是用所有实验者的部分样本参与模型训练，比只用部分实验者的全部样本参与模型训练，训练出来的模型更为鲁棒。

3 结论

本文提出一种单视角下基于投影子空间视图的动作识别方法。具体采用3D动作的三视图将高维空间中的动作识别问题转化为多个低维子空间的动作识别问题，并运用投票策略融合三个视图的动作信息输出最终识别结果。实验结果表明，本文方法在一定程度上解决了单视角下观察动作时存在的关节点自遮挡问题，且达到了与多视角动作数据集相当的识别效果。但是本文算法也存在一定的局限性，例如未对3D动作三视图的重要性加以区分，而实际上，不同视图观察动作获得的信息对重构3D动作发挥作用的大小不同；其次是本文仅使用关节点坐标数据进行特征学习与动作识别，而关节点坐标数据相较于关节距离和关节角是低阶特征，对动作的表征能力有限。

未来研究工作，将会对3D动作三视图对动作识别的重要性加以区分，即赋予主视图、左视图、俯视图不同的权重进行加权融合；在输入端丰富动作的表征信息，即同时输入骨骼数据的点级和边级数据。具体地，在骨骼数据的基础上引入关节距离和关节角特征，三种动作特征分别结合ST-GCN并行进行动作识别；根据动作骨骼数据反算捕获数据的相机方位，并以此确定动作子视图的权重用于动作识别。

参考文献：

[1] Martinez-Hernandez U, Dodd T J, Prescott T J. Feeling the Shape: Active Exploration Behaviors for Object Recognition with a Robotic Hand[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems (S2168-2216), 2018, 48(12): 2339-2348.

[2] 吴培良, 杨霄, 毛秉毅, 等. 一种视角无关的时空关联深度视频行为识别方法[J]. 电子与信息学报, 2019, 41(4): 904-910.
Wu Peiliang, Yang Xiao, Mao Bingyi, et al. A Perspective-Independent Method for Behavior Recognition in Depth Video Via Temporal-Spatial Correlating[J]. Journal of Electronics & Information Technology, 2019, 41(4): 904-910.

[3] Sudha M R, Sriraghav K, Sudar Abisheck S, et al. Approaches and Applications of Virtual Reality and Gesture Recognition: A Review[M]. Hershey, PA, USA: IGI Global, 2018: 180-199.

[4] 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.

[5] Johansson G. Visual Perception of Biological Motion and A Model for Its Analysis[J]. Perception & Psychophysics (S1532-5962), 1973, 14(2): 201-211.

[6] Yan Sijie, Xiong Yuanjun, Lin Dahua. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[C]// Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2018: 7444-7452.

[7] Shi Lei, Zhang Yifan, Cheng Jian, et al. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 12018-12027.

[8] Qin Zhenyue, Liu Yang, Ji Pan, et al. Fusing Higher-Order Features in Graph Neural Networks for Skeleton-Based Action Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. arXiv: 2105.01563[cs.CV].

[9] Sun Dengdi, Zeng Fanchen, Luo Bin, et al. Information Enhanced Graph Convolutional Networks for Skeleton-Based Action Recognition[C]// 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway, NJ, USA: IEEE, 2020: 1-7.

[10] Feng Dong, Wu Zhongcheng, Zhang Jun, et al. Multi-scale Spatial Temporal Graph Neural Network for Skeleton-Based Action Recognition[J]. IEEE Access

(S2169-3536), 2021, 9: 58256-58265.

[11] Shi Lei, Zhang Yifan, Cheng Jian, et al. Skeleton-Based Action Recognition with Directed Graph Neural Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 7904-7913.

[12] Ji Xiaofei, Liu Honghai. Advances in View-Invariant Human Motion Analysis: A Review[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (S1094-6977), 2010, 40(1): 13-24.

[13] Su Benyue, Wu Huang, Sheng Min, et al. Accurate Hierarchical Human Actions Recognition from Kinect Skeleton Data[J]. IEEE Access (S2169-3536), 2019, 7: 52532-52541.

[14] Cao Congqi, Lan Cuiling, Zhang Yifan, et al. Skeleton-Based Action Recognition with Gated Convolutional Neural Networks[J]. IEEE Transactions on Circuits and Systems for Video Technology (S1051-8215), 2019, 29 (11): 3247-3257.

[15] Li Bo, Dai Yuchao, Cheng Xuelian, et al. Skeleton Based Action Recognition Using Translation-Scale Invariant Image Mapping and Multi-scale Deep CNN[C]// 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway, NJ, USA: IEEE, 2017: 601-604.

[16] 付仔蓉, 吴胜昔, 吴潇颖, 等. 基于空间特征的BI-LSTM人体行为识别[J]. 华东理工大学学报(自然科学版), 2021, 47(2): 225-232.
Fu Zairong, Wu Shengxi, Wu Xiaoying, et al. Human Action Recognition Using BI-LSTM Network Based on Spatial Features[J]. Journal of East China University of Science and Technology, 2021, 47(2): 225-232.

[17] 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 (CVPR). Piscataway, NJ, USA: IEEE, 2016: 1010-1019.

[18] Liu Jun, Shahroudy A, Perez M, et al. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (S0162-8828), 2020, 42(10): 2684-2701.

[19] Ke Qiuhong, Bennamoun M, An Senjian, et al. Learning Clip Representations for Skeleton-Based 3D Action Recognition[J]. IEEE Transactions on Image Processing (S1057-7149), 2018, 27(6): 2842-2855.

[20] 徐飞翔. 基于视角估计的自由视点动作理解[D]. 成都: 电子科技大学, 2021: 19-37.
Xu Feixiang. The Analysis of Action Understanding Under Arbitrary-View Based on the Viewpoint Estimation [D]. Chengdu: University of Electronic Science and Technology of China, 2021: 19-37.

[21] 孙鹏辉. 基于深度学习的多视图三维重建算法及并行化研究[D]. 银川:宁夏大学, 2020: 1-7.
Sun Penghui. Multi-view 3D Reconstruction Algorithm and Parallelization Research Based on Deep Learning [D]. Yinchuan: Ningxia University, 2020: 1-7.

[22] 胡文涛, 陈秀宏. 基于局部保持投影的鲁棒稀疏子空间学习[J]. 计算机工程与应用, 2021, 57(10): 194-199.
Hu Wentao, Chen Xiuhong. Robust Sparse Subspace Learning Based on Locality Preserving Projections[J]. Computer Engineering and Applications, 2021, 57(10): 194-199.

[23] 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). Piscataway, NJ, USA: IEEE, 2017: 1623-1631.

[24] Ke Qiuhong, Bennamoun M, An Senjian, et al. A New Representation of Skeleton Sequences for 3D Action Recognition[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 4570-4579.

[25] Liu Jun, Shahroudy A, Xu Dong, et al. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (S1939-3539), 2018, 40(12): 3007-3021.

[26] Li Chaolong, Cui Zhen, Zheng Wenming, et al. Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition[C]// AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2018: 3482-3489.

[27] Liu Jun, Wang Gang, Duan Lingyu, et al. Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks[J]. IEEE Transactions on Image Processing (S1941-0042), 2018, 27(4): 1586-1599.

[28] Yang Zhengyuan, Li Yuncheng, Yang Jianchao, et al. Action Recognition with Spatio-Temporal Visual Attention on Skeleton Image Sequences[J]. IEEE Transactions on Circuits and Systems for Video Technology (S1558-2205), 2019, 29(8): 2405-2415.

[29] Qin Yang, Mo Lingfei, Li Chenyang, et al. Skeleton-Based Action Recognition by Part-Aware Graph Convolutional Networks[J]. The Visual Computer (S0178-2789), 2020, 36(3): 621-631.

