文章编号：1007-2780(2022)12-1614-12

基于多尺度多分支特征的动作识别

张　磊[1,2]，韩广良[1*]
（1. 中国科学院 长春光学精密机械与物理研究所，吉林 长春 130033；
2. 中国科学院大学，北京 100049）

摘要：针对基于人体骨架序列的动作识别存在的特征提取不充分、不全面及识别准确率不高的问题，本文提出了基于多分支特征和多尺度时空特征的动作识别模型。首先，利用多种算法的结合对原始数据进行了特征增强；其次，将多分支的特征输入形式改进为多分支的融合特征信息并分别输入到网络中，经过一定深度的网络模块后融合在一起；然后，构建多尺度的时空卷积模块作为网络的基本模块，用来提取多尺度的时空特征；最后，构建整体网络模型输出动作类别。实验结果表明，在 NTU RGB-D 60 数据集的两种划分标准 Cross-subject 和 Cross-view 上的识别准确率分别为 89.6% 和 95.1%，在 NTU RGB-D 120 数据集的两种划分标准 Cross-subject 和 Cross-setup 上的识别准确率分别为 84.1% 和 86.0%。与其他算法相对比，本文算法提取到了更为多样化、多尺度的动作特征，动作类别的识别准确率有一定的提升。

关　键　词：动作识别；多尺度特征；多分支特征；特征融合
中图分类号：TP391.4　　**文献标识码**：A　　**doi**：10.37188/CJLCD.2022-0176

Action recognition algorithm based on multi-scale and multi-branch features

ZHANG Lei[1,2], HAN Guang-liang[1*]

(1. *Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Acadamy of Sciences, Changchun* 130033, *China*;
2. *University of Chinese Acadamy of Sciences, Beijing* 100049, *China*)

Abstract: Aiming at the problems of insufficient feature extraction, incompleteness and low recognition accuracy in action recognition based on human skeleton sequence, a action recognition model based on multi-branch feature and multi-scale spatio-temporal feature is proposed in this paper. Firstly, the original data are enhanced by the combination of various algorithms. Secondly, the multi-branch feature input form is improved to multi-branch fusion feature information, which is input into the network, respectively. After a certain depth of network modules, it is fused together. Then, a multi-scale spatio-temporal convolution module is constructed as the basic module of the network to extract multi-scale spatio-temporal features. Finally, the overall network model is constructed to output action categories. The experimental results show that the recognition accuracy on Cross-subject and Cross-view of NTU RGB-D 60 data set is 89.6% and

收稿日期：2022-05-25；**修订日期**：2022-06-06.
基金项目：吉林省科技厅重点项目（No. 20210201132GX）
Supported by Department of Science and Technology of Jilin Province (No. 20210201132GX)
*通信联系人，E-mail: hangl@ciomp.ac.cn

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

95.1%, and the recognition accuracy on Cross-subject and Cross-setup of NTU RGB-D 120 data set is 84.1% and 86.0%, respectively. Compared with other algorithms, the more diversified and multi-scale action features are extracted, and the recognition accuracy of action categories is improved to a certain extent.
Key words: action recognition; multi-scale features; multi-branch features; feature fusion

1 引　　言

人体动作识别是指根据输入的动作时间序列(RGB图像序列或人体骨骼点序列)进行分析，从而识别出该动作序列所进行的动作类别，是近年来的研究热点之一。人体动作识别在智能监控系统、人机交互、动作辅助矫正等诸多领域有着广泛的应用前景。同时由于人体骨架序列的数据量极小，远小于RGB视频序列，且结构简单、计算速度快，几乎不会受到背景的影响，所以非常适合于动作识别的应用。

在早期的研究中，研究人员常通过人为设计多种动作特征来代表的人体的运动轨迹。Dalal等首先提出使用光流直方图的方法来获取运动轨迹[1]，随后Dalal等又提出了方向梯度直方图的方法[2]。在此基础上，Wang等提出了密集轨迹法(Dense Trajectory，DT)及其改进方法增强的密集轨迹法(improved Dense Trajectory，iDT)，通过大量采样视频中随机的像素点追踪时间维度的该像素点的运动轨迹[3-4]。由于人工设计的特征工作量较大、准确率低，所以基于神经网络的方法逐渐应用于动作识别中[5]。Du等于2015年首次提出了将卷积神经网络(Convolutional Neural Network，CNN)应用于骨骼动作识别的SK-CNN，将人体骨架序列结构归一化为固定大小输入到CNN网络中进行动作识别[6]。由于人体骨架序列与CNN的输入数据结构存在差异，Wang等提出了关节轨迹图(Joint Trajectory Maps，JTM)，将骨架数据投影到3个正交平面从而得到3幅2D骨架数据，归一化后送入预训练好的CNN进行动作识别[7]。Li等借鉴共现特征学习的方法提出了HCN网络[8]。基于CNN的方法往往参数量较大，也存在遮挡、视点变化等问题，所以基于循环神经网络(Recurrent Neural Network，RNN)和基于图卷积网络(Graph Convolution Neural Network，GCN)的算法[9-10]也开始逐渐出现。Shahroudy等提出利用长短时记忆网络(Long Short-Term Memory，LSTM)进行动作识别，但无法获得具有竞争力的结果[11]。由于图卷积网络(Graph Convolution Neural Network，GCN)对具有拓扑结构的人体骨架数据处理十分有效，所以Yan等首先在GCN的基础上提出了ST-GCN，利用图卷积提取人体动作的时空特征[12]。Li等在ST-GCN的基础上提出AS-GCN，揭示了关节点之间的潜在关系[13]。

基于人体骨架序列的动作识别相比于基于RGB视频的方法具有数据量小，不受外界环境、光照、人体外貌等因素影响，鲁棒性高的优点，成为当下的研究热点。而在基于骨架的方法中，基于CNN的方法是主流，也取得了较好的成果，但其网络大多依赖于规模较大的网络，计算量较大，且难以直接处理人体骨架序列这样的拓扑结构。基于RNN的方法一般识别准确率不如其他方法，并且随着神经网络层数的不断加深，极易出现梯度消失的现象，使得训练的准确率突然降低。基于GCN的方法具有CNN的优点，同时由于引入了人体拓扑结构的先验知识，能够进一步提升识别性能。除此之外，现阶段的算法尚不能在多种运动特征之间发现更深层次的关系，并且对于提取到的多种运动特征信息都很难得到高效的应用。

针对以上问题，本文首先根据动作识别网络模型的多分支输入和原始数据的结构，对原始数据进行了特征增强，将多分支的输入形式改进为多分支的融合特征，并经过多个网络模块后融合多层次多种类的特征信息，从而增强了多特征之间的相关性，从特征层面更全面地描述一个人体骨架序列。其次，本文通过多尺度特征对时序卷积进行了改进，并与图卷积神经网络结合，使得网络模型能够提取不同深度的时间特征和空间特征，提高了动作识别的准确性。通过残差连接图卷积神经网络和多尺度时序卷积，在提取到深层次特征信息的同时很大程度上解决了网络退化等问题。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2 算法的构成及原理

2.1 原始数据的特征增强

原生的人体骨架序列数据在采集时可能存在数值缺失、数值范围小等问题，同时由于设备精度问题使得骨架序列在时间维度上存在与动作无关的抖动，并且在不同的动作类别中，骨架序列的帧数也不同。针对以上问题，本文分别在时间和空间两个维度进行如图1所示的特征增强。归一化、坐标转换、深度优先树遍历等基于空间维度的特征增强重点关注于关节点的位置坐标和联系上，Savitsky-Golay平滑滤波、插帧等基于时间维度的特征增强重点关注于骨架序列帧间的联系。

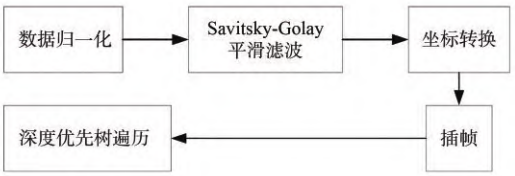

图1 特征增强处理方法

Fig. 1 Feature enhancement processing methods

数据集中的关节点坐标的大小常位于小数点后一位或两位，使得动作的关节点过于密集，不利于动作类别的预测。最好的处理方法是将整体坐标所处的区间放大，所以采用归一化的方式将关节点坐标的区间扩大至(−1,1)，新坐标可表示为：

$$J'_{t,c}=\frac{2\times\left(J_{t,c}-J_{t,c,\min}\right)}{J_{t,c,\max}-J_{t,c,\min}}-1\ , \tag{1}$$

其中$J_{t,c}$、$J_{t,c,\max}$、$J_{t,c,\min}$分别表示在某一动作中t时刻x、y、z分量上的坐标。

原始的人体骨架数据沿着时间维度运动时会存在一些与动作无关但影响类别判断的噪音，这一噪音是人体骨架数据在可视化时运动中关节点的不自然抖动。针对这一问题，本文使用Savitsky-Golay滤波器进行滤波。Savitsky-Golay滤波器是一种在时域内基于局域多项式最小二乘法拟合的滤波方法，这种滤波器最大的特点在于在滤除噪声的同时可以确保信号的形状、宽度不变。Savitsky-Golay滤波器通过多项式对移动窗口内的数据进行多项式最小二乘拟合，算出窗口内中心点关于其周围点的加权平均和，可表示为：

$$X_i^*=\frac{\sum_{j=-r}^{r}X_{i+j}W_j}{\sum_{j=-r}^{r}W_j}\ , \tag{2}$$

其中X_i和X_i^*为滤波前、后的数据，W_j为移动窗口平滑过程中的权重因子，窗口长度为$2r+1$。窗口长度是平滑滤波中最重要的参数，若窗口长度过小则噪音无法减弱，若窗口长度过长则会将原本正常的动作变化幅度减弱甚至消减至静止，取$r=4$。

接着为了确保关节点的参考坐标系一致，同时减少采集过程中设置的相机多视角产生的影响，以图2所示的骨架结构及编号为基础，将每个动作序列中的第一帧骨架的脊柱中心点(编号20)设置为参考坐标系的原点，脊柱(编号0和1)所在的直线作为z轴，两侧肩膀(编号4和8)所在的直线作为x轴。原始的骨架结构数据中各个类别的动作帧数不是统一的，多则接近300帧，少则少于50帧，无法直接输入到后续网络中。针对这一问题，本文采用插帧的方式，将帧数设置为300，采用三次样条插值法进行处理。若区间$[a,b]$可分为n个区间$[(x_0,x_1),(x_1,x_2),(x_2,x_3),\cdots,(x_{n-1},x_n),]$，其中$a=x_0$，$b=x_n$，在$n$个区间中的每个区间都各自存在三次样条函数$S_i(x)$，可表示为：

$$y=a_i+b_ix+c_ix^2+d_ix^3\ , \tag{3}$$

其中a_i、b_i、c_i和d_i为$S_i(x)$在n个区间中的第i个小区间的参数。$S_i(x)$必须满足3个条件：在每个

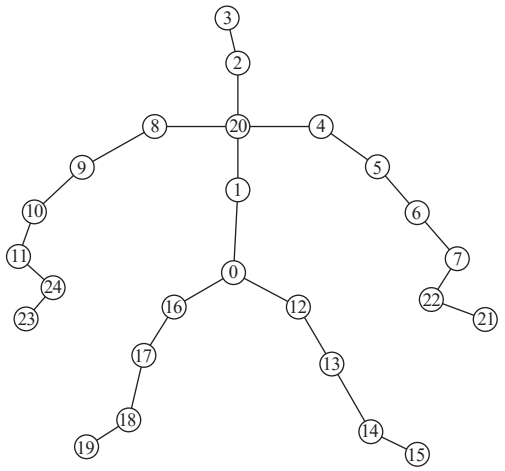

图2 人体骨架结构及编号

Fig. 2 Human skeleton structure and number

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

小区间(x_{n-1},x_n)内$S_i(x)$都是一个三次方程；满足插值条件，即x_n必须在$S(x)$函数的曲线上；$S_i(x)$的曲线是光滑的，即$S(x)$、$S'(x)$、$S''(x)$是连续的。可计算得到：

$$a_i=y_i\ , \tag{4}$$

$$b_i=\frac{y_{i+1}-y_i}{h_i}-\frac{h_i m_i}{2}-\frac{h_i}{6}(m_{i+1}-m_i)\ , \tag{5}$$

$$c_i=\frac{m_i}{2}\ , \tag{6}$$

$$d_i=\frac{m_{i+1}-m_i}{6h_i}\ , \tag{7}$$

其中$m_i=S''_i(x_i)$，$h_i=x_{i+1}-x_i$。

为了进一步有效学习人体骨架中关节点的空间关系，引入了深度优先树遍历[14]。输入格式中一对相邻关节的空间相关性与连接这一对关节的边有关，若这对关节在物理上是相连的，则空间相关性就强。在动作识别中，空间相关性一般是指人体骨骼结构中相邻关节点之间的联系紧密程度。空间相关性越强，最后所反映的动作识别准确率越高。原始数据的输入形式是将抽象的关节点坐标序列转化为$T\times N\times C$的矩阵，其中T为人体动作序列的帧数，N为人体骨架的关节点数量，C为关节点坐标的维度数。在代表关节点数量的N维度中，人体的25个关节是以0,1,2,3,…,23,24这样顺序编号进行排列的，如图3(a)所示，相邻的关节在图2所示的结构中基本上是不相连的，关节之间便没有相关性。根据这一理论，可将人体骨架结构改变为图4所示的树形骨架结构图。根据深度遍历，按照从上到下、从左到右的顺序，在输入数据的N维度上，人体的关节可排序为1,20,2,3,2,20,4,5,6,7,22,21,22,7,6,5,4,20,8,9,10,11,24,23,24,11,10,9,8,20,1,0,12,13,14,15,14,13,12,0,16,17,18,19,18,17,16,0,1,N由原来的25变为49。

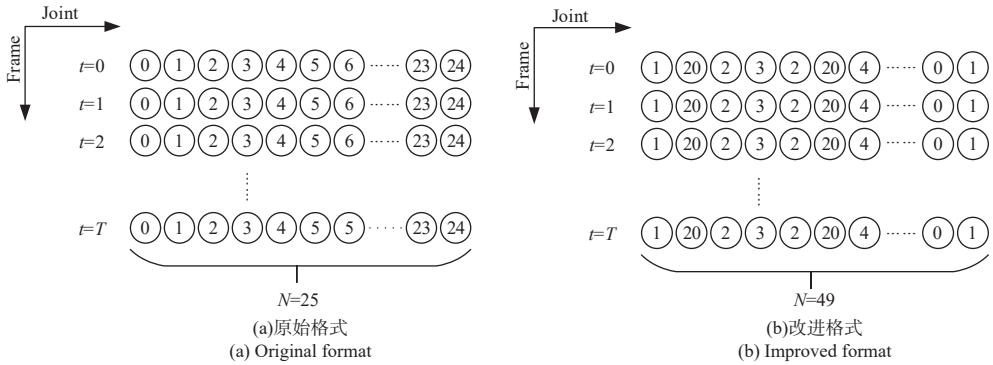

图3　数据输入的原始和改进格式

Fig. 3　Original and improved formats for input

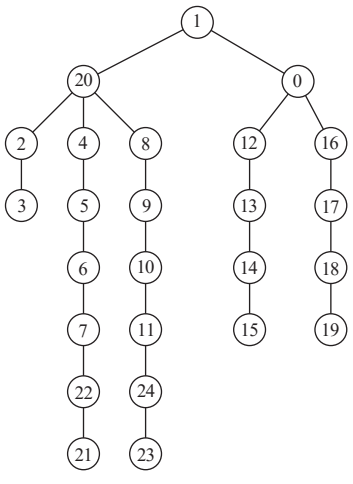

图4　树形骨架结构图

Fig. 4　Structure diagram of tree structure skeleton

2.2　多分支输入

经过特征增强后，单一样本中的人体骨架数据结构可表示为四维向量(C,T,V,M)，其中C表示骨架关节点的x、y、z分量；T代表骨架数据的帧数，经过特征增强后固定为300帧；V代表每一帧中的人体关节点数量；M代表每一个动作的参与人数。参考ResGCN[15]中关节点坐标(Joints)、运动向量(Velocities)和骨节(Bones)三分支输入，为了增加输入数据的空间相关性，本文将三分支输入修改为Bone length＋Joints、Bone length＋Velocities和Bone length＋Bone angle。

多分支输入的操作需要将特征融合在一起，从而将经过一定深度的神经网络提取到的特征

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

拼接在一起，此时的特征在经过多层神经网络后具有一定的深层信息，而特征融合的具体位置在哪一层神经网络之后则需要进一步的实验与研究。

2.3 多尺度时空卷积

本文网络模型的基本结构由可提取时间和空间特征的多尺度时空卷积构成。多尺度指的是对网络中不同深度的特征进行采集融合。多尺度时空卷积模块由负责提取时间特征的多尺度时序卷积和负责提取空间特征的图卷积组成，使得网络能够分别提取网络中的时间特征和空间特征。

本文中的多尺度时序卷积模块如图5所示。网络模块中的第一层是卷积核为1×1的普通卷积，其作用是降低输入数据的通道维数；模块中的第二层为膨胀系数分别为1，2，3的膨胀卷积[16]，卷积核为$m\times1$。原本大小为$m\times n$的卷积核要对两个维度的特征进行处理，但当$n=1$时，卷积核只会处理一个维度的特征，因此通过这种方式来只对时间维度进行特征提取。多尺度特征通过设置不同的膨胀系数来体现，当膨胀卷积拥有不同的膨胀系数时，卷积核的感受野也随之变化，卷积所能提取到的特征也会有所区分。膨胀卷积的优点在于在保持参数个数不变的情况下增大了卷积核的感受野，让每个卷积输出都包含较大范围的信息，同时它可以保证输出的特征映射的大小保持不变。

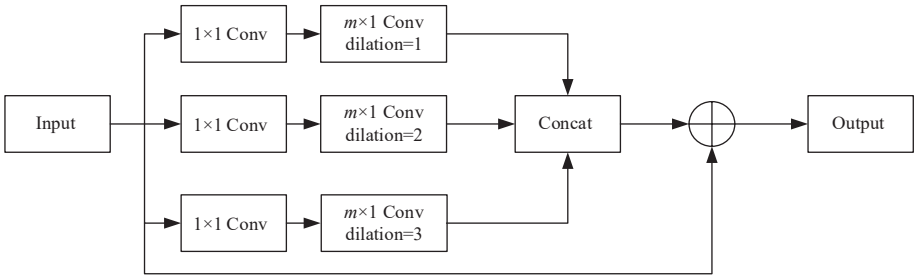

图5 多尺度时序卷积模块结构图

Fig. 5 Structure diagram of multi-scale temporal convolution

输入特征在经过多个并列的膨胀卷积提取多尺度时间特征后，使用Concat函数进行拼接，得到的特征通道维度与输入一致，此时使用残差将输入特征与输出特征相连接，使得输出同时具有浅层网络的特征表现和深层网络的特征表现，有助于解决梯度消失和梯度爆炸问题，在网络层数不断加深的同时，又能保证良好的性能。

GCN处理的主要对象是骨架数据、交通网络数据、化学分子结构数据等具有一定拓扑结构的数据。在动作骨架的每一帧中，图卷积可表示为

$$f_{out}=\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{\Lambda}^{-\frac{1}{2}}f_{in}\boldsymbol{W}\ , \tag{8}$$

其中f_{in}和f_{out}分别代表输入和输出特征，$\boldsymbol{A}$表示邻接矩阵，$\boldsymbol{I}$为单位矩阵，$\boldsymbol{W}$为权重矩阵，$\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{\Lambda}^{-\frac{1}{2}}$表示归一化邻接矩阵。

本文图卷积的邻接矩阵使用距离分区(Distance partitioning)的方法。对于一个庞大的图结构而言，不可能把所有的节点特征直接相加。为了划分图结构中每一个节点的邻居节点，即骨骼结构中节点附近的节点，采用距离分区的方法将不同的邻居节点编为不同的序号，如图6所示，同一个序号的邻居节点看作一个邻居子集，也就形成了多个邻接矩阵。本文中分区子集设为4个，即距离目标节点的距离分别为0，1，2，3。距离为0

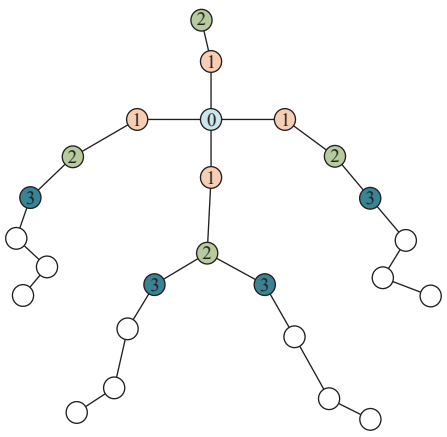

图6 距离分区的划分原理

Fig. 6 Division principle of distance partitioning

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

为节点自身，位于邻接矩阵的一个子集中；距离为1的节点为与目标节点直接相连，位于邻接矩阵的第二个子集中；距离为2和3的节点则分别位于邻接矩阵的第三和第四个子集中。

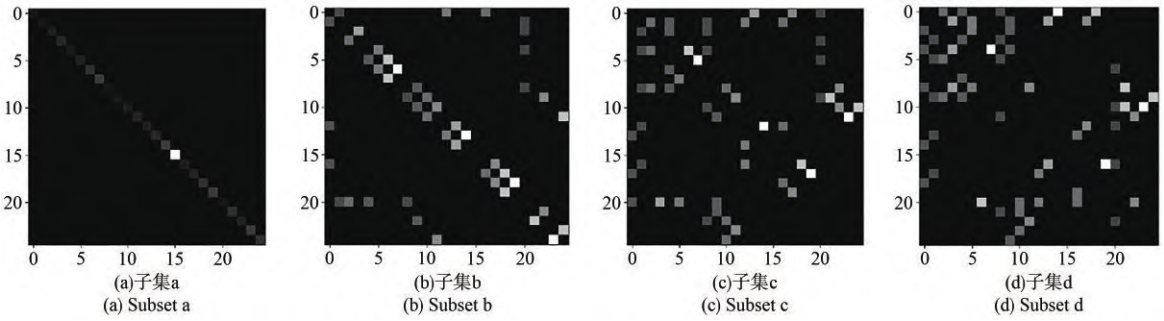

图7　邻接矩阵可视化结果

Fig. 7　Visualization results of adjacency matrix

由此得到的邻接矩阵可视化结果如图7所示。使用了距离分区的分区策略后，图卷积可表示为：

$$f_{\text{out}}=\sum_j \boldsymbol{\Lambda}_j^{-\frac{1}{2}}\left(\boldsymbol{A}_j\otimes\boldsymbol{M}\right)\boldsymbol{\Lambda}_j^{-\frac{1}{2}}f_{\text{in}}\boldsymbol{W}_j\,, \tag{9}$$

其中$j=0,1,2,3$，表示邻接矩阵的子集；$\boldsymbol{M}$为权重矩阵，使用了距离分区的邻接矩阵依照距离的大小为子集划分不同的权重。基于距离分区的分区策略能够建模如关节之间的相对平移等关节点的局部差异性，增加图卷积提取空间特征的多样性。

多尺度时空卷积模块的结构如图8所示。在一定程度上，网络的层数越深则表达能力越强，越能够提取到深层次的特征信息，性能也越好。但随着网络深度的增加，也带来了许多问题，如梯度消散、网络退化等，残差连接可以在很大程度上解决这一问题。

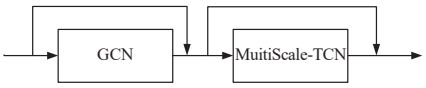

图8　多尺度时空卷积模块结构图

Fig. 8　Structure image of multi-scale spatial-temporal convolution

2.4　整体网络结构搭建

整体网络结构如图9所示，由多尺度时空卷积模块堆叠而成。原始的人体骨架序列存在诸多缺陷，特征提取阶段通过基于时间维度和空间维度的特征增强加强了数据的表达能力。将处理好的数据以Bone length＋Joints、Bone length＋Velocities和Bone length＋Bone angle的融合特征的方式分别输入到三分支中，利用改进的多尺度卷积提取特征并融合在一起。动作分类阶段将拼合后的数据输入带有注意力机制的多尺度卷积进行多次训练，最后进入分类器进行动作分类。

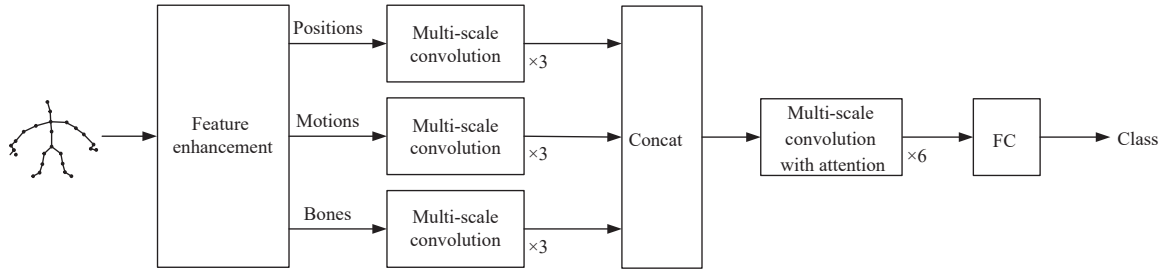

图9　网络结构示意图

Fig. 9　Schematic diagram of network structure

为了更加全面地认识本文所设计的网络结构，本节详细给出了多尺度时空卷积模块的结构和在每一层所使用的参数，如表1所示。表1中Conv_gcn代表图卷积模块，Conv_tcn代表多尺度

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表1 多尺度卷积网络参数表

Tab. 1 Network parameter table of multi-scale convolution

Number of layers	Module name	Size	Step	Input and output
1	Conv1	1×1	1	6, 36
2	Conv_gcn	3×3	1	36, 72
	Conv_tcn	1×1	1	72, 24
		5×1	2	24, 24
3	Conv_gcn	3×3	1	72, 72
	Conv_tcn	1×1	1	72, 24
		5×1	1	24, 24
4	Conv_gcn	3×3	1	72, 36
	Conv_tcn	1×1	1	36, 12
		5×1	1	12, 12
5	Conv_gcn	3×3	1	108, 192
	Conv_tcn	1×1	1	192, 64
		5×1	2	64, 64
6,7	Conv_gcn	3×3	1	192, 192
	Conv_tcn	1×1	1	192, 64
		5×1	2	64, 64
8	Conv_gcn	3×3	1	192, 384
	Conv_tcn	1×1	1	384, 128
		5×1	1	128, 128
9,10	Conv_gcn	3×3	1	384, 384
	Conv_tcn	1×1	1	384, 128
		5×1	2	128, 128
11	FC	—	—	384, 60

时间卷积模块,包含多个不同尺度的卷积。

3 实验结果与分析

3.1 实验环境

本实验使用的数据集为NTU RGB-D 60和NTU RGB-D 120的人体骨架结构数据。在RTX 3090上进行训练,显存为25.4 GB。所利用的环境为Python 3.6,PyTorch 1.10.0。

在训练过程中,所有需要学习的参数均用Xavier方法进行初始化。网络模型使用交叉熵损失函数(Cross Entropy)进行训练,优化器采用随机梯度下降法(Stochastic Gradient Descent,SGD)对网络参数进行优化,学习率为0.1,数据集在迭代70次(Epoch)后停止训练,批(Batch)大小设置为16。

3.2 特征增强方法对比实验分析

为了充分证明每种特征增强方法对于动作识别的作用,需要分别对每一种方法进行实验测试。基于NTU RGB+D 60数据集的Cross-view(CV)划分标准,将经过各自方法进行特征增强后的数据输入到网络中,得到的识别准确率如表2所示。

由实验结果可以看出,每种方法都取得了一定的识别准确率。在这5种特征增强方法单独的实验结果中,只使用深度优先树遍历的模型得到了最高的识别准确率,说明基于深度优先树遍历的方法的有效性得到了实验验证。其次识别准确率最高的是基于坐标转换的动作识别,接着3种普通的数据增强方法排在后面,且与前两种的准确率差别较大,说明网络对于从多个层次描述空间特征是十分认可的。将所有方法一起使用的动作识别准确率为95.1%,与单独的特征增强方法得到的识别准确率相差至少2.5%,说明本文使用的5种特征增强方法都或多或少发挥了其应有的作用,对于人体骨架动作的识别都有正向的作用,为后续网络提取到多层次、多尺度的时间和空间特征打下了坚实基础。

表2 不同特征增强方法在NTU RGB-D 60数据集的准确率

Tab. 2 Accuracy of different feature enhancement method on NTU RGB-D 60 dataset

	数据归一化/%	SavGol滤波/%	坐标转换/%	插帧/%	深度优先树遍历/%	应用所有方法的特征增强/%
Cross-view (CV)	89.3	90.1	92.1	90.5	92.6	95.1

3.3 多分支输入方法对比实验分析

本文改善了三分支输入的人体骨架特征,使得三分支的输入变为Bone length+ Joints、Bone length+ Velocities和Bone length+ Bone angle,增加了输入特征的关节相关性。为了验证每一分支对于动作类别判断的有效性,分别将3种特

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

征输入网络中进行实验，在NTU RGB+D 60数据集的Cross-view（CV）评价标准上的识别准确率如表3所示。

由实验结果可知，3种人体骨架结构特征都在网络中取得了相接近的动作识别准确率。在3种特征Bone length+Joints、Bone length+Velocities和Bone length+Bone angle分别得到的动作识别准确率中，特征Bone length+Joints的识别准确率最高，而特征Bone length+Velocities和特征Bone length+Bone angle紧随其后，二者之间的识别准确率差别不大，仅在0.2%，说明3种特征均有一定的能力独自作为网络的输入，可以从中提取到一定程度的信息用以动作分类。当三分支同时输入3种特征时，其结果为95.1%，与3种特征分别输入得到的结果相比，有了一定程度的提升，说明3种特征以三分支的形式输入到网络中对动作识别的准确率起到积极的促进作用。

表3　不同特征在NTU RGB-D 60数据集的准确率

Tab. 3　Accuracy of different feature on NTU RGB-D 60 dataset

	Bone length+Joints/%	Bone length+Velocities/%	Bone length+Bone angle/%	同时输入三种特征/%
CV	93.8	93.0	92.8	95.1

3.4　多分支输入方法对比实验分析

本小节主要探讨三分支特征在特征融合时的位置对动作类别判断产生的影响。若特征融合时的位置在第k个多尺度卷积模块的后方，则取k=1，2，3，4时网络的识别准确率如表4所示。

表4　不同融合位置在NTU RGB-D 60数据集的准确率

Tab. 4　Accuracy of different feature fusion location on NTU RGB-D 60 dataset

	k=1	k=2	k=3	k=4
CV	92.8%	94.5%	95.1%	94.1%

由消融实验的结果可知，不同的k值取得了不同的动作识别准确率。当k值为3时，人体动作的识别准确率最高，所以本文中特征融合的阶段选择在第三个多尺度卷积模块之后。在本实验中，当k值过小时，多分支的特征融合后虽然具有很强的浅层次相关性，但在融合后经过神经网络层数的不断深入，提取到的深层次特征趋为一致，缺少了多样化的深层次特征；而当k值过大时又缺少了早期特征之间的相关性。所以，要选择合适的k值。

实验结果表明，特征融合的位置会对动作识别的准确率产生影响，应该选择能够使准确率达到最大的位置。

3.5　多尺度时序卷积结构对比实验分析

本小节主要探讨多尺度时序卷积的卷积核大小对人体动作识别性能的影响。使用NTU RGB-D 60数据集的Cross-view（CV）划分标准作为消融实验的数据集，时序卷积的卷积核m=3，5，7，9时的动作识别准确率如表5所示。

表5　不同时序卷积核在NTU RGB-D 60数据集的准确率

Tab. 5　Accuracy of different temporal convolution kernel on NTU RGB-D 60 dataset

	m=3	m=5	m=7	m=9
CV	94.5%	95.1%	94.7%	93.4%

多尺度时序卷积的卷积核是提取时间维度特征的核心。卷积核越大，特征信息的感受野也随之变大，多尺度时序卷积能够提取到的时间维度的跨度就越大，也就是说能够提取距离更长的视频帧间的特征信息，但同时也容易漏掉更多的细节信息，所以选择合适的卷积核大小尤为重要。由实验结果可知，当卷积核的大小m=5时，人体动作识别在RGB-D 60数据集Cross-view（CV）的准确率最高，为95.1%。当卷积核过小时，提取到的特征在时间维度的跨度不够长，帧间特征无法高效地表达动作，同时网络的参数量和计算量也相较较大；当卷积核过大时，容易漏掉更多的细节信息，帧间特征无法完整地表达动作。所以多尺度时序卷积的卷积核大小设为5×1。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3.6 对比实验

使用NTU RGB-D 60数据集与NTU RGB-D 120数据集展示本文提出的网络模型的结果，并与目前提出的其他基于人体骨架数据的动作识别算法进行对比与分析，在NTU RGB-D 60数据集上的人体动作识别准确率如表6所示。

表6 不同动作识别方法在NTU RGB-D 60数据集的准确率

Tab. 6 Accuracy of different action recognition methods on NTU RGB-D 60 dataset

Method	Cross-subject/%	Cross-view/%
TCN[17]	74.3	83.1
Ind-RNN[18]	81.8	88.0
ST-GCN[12]	81.5	88.3
HCN[8]	86.5	91.1
AS-GCN[13]	86.8	94.2
3SCNN[19]	88.6	93.7
SGN[20]	86.6	94.3
PR-GCN[21]	85.2	91.7
PeGCN[22]	85.6	93.4
Ours	89.6	95.1

本文提出的网络模型在NTU RGB-D 60数据集的Cross-subject划分标准上获得了89.6%的识别准确率，在Cross-view划分标准上获得了95.1%的准确率。与基于循环神经网络的Ind-RNN网络模型相比，在两个指标上分别高出7.8%和7.1%；与基于卷积神经网络的3SCNN网络模型相比，在两个指标上分别高出1.0%和1.4%；与同样基于图卷积神经网络的SGN相比，在两个指标上分别高出3.0%和0.8%。与其他基于人体骨架数据的动作识别算法相对比的结果说明了本文所使用网络的优越性。

横向对比基于NTU RGB-D 60数据集的Cross-subject和Cross-view划分标准的识别准确率结果，可以发现所有识别准确率在Cross-view评价标准上的结果均要高于Cross-subject评价标准，两者之差最大能够达到8.8%，最小能够达到4.6%。

为了对实验结果进行更深入的分析，得到图10所示的Cross-subject和Cross-view划分标准下的混淆矩阵图，图中横坐标为具体动作类别的编号，纵坐标为具体动作类别的准确率。由图10可知，模型在Cross-subject中可准确识别绝大部分的动作类别。其中类别11(Reading)、12(Writing)、29(Play with phone/Tablet)、30(Type on a keyboard)、34(Rub two hands)、41(Sneeze/Cough)、44(Headache)的准确率明显低于其他类别。这些动作类别聚焦于局部肢体动作，动作幅度不大，动作之间的区别也不明显。模型无法确切区分这些动作类别，这是影响本文模型准确率的主要因素之一。

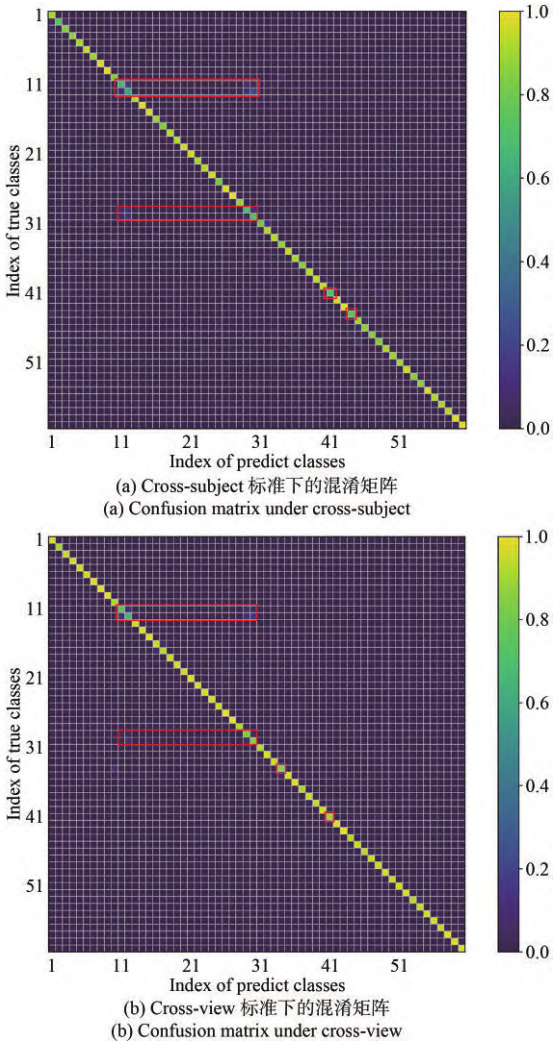

图10 NTU RGB-D 60数据集下的混淆矩阵

Fig. 10 Confusion matrix under NTU RGB-D 60 dataset

除了NTU RGB-D 60数据集之外，本文也在其扩展版本NTU RGB-D 120数据集上进行了实验与分析，如表7所示。该数据集相比于NTU RGB-D 60数据集，其动作类别数更多且人物、场景情况更复杂，因此基本上所有主流方法在该数据集上的识别准确率均出现了明显的下降。在NTU RGB-D 120数据集中，本文提出的网络模型

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 7　不同动作识别方法在NTU RGB-D 120数据集的准确率

Tab. 7　Accuracy of different action recognition methods on NTU RGB-D 120 dataset

Method	Cross-subject/%	Cross-setup/%
Part-aware LSTM[11]	25.5	26.3
TSRJI[23]	67.9	62.8
3SCNN[19]	81.2	81.9
3s RA-GCN[24]	81.1	82.7
Gimme Signals[25]	70.8	71.6
SGN[20]	79.2	81.5
Ours	84.1	86.0

在Cross-subject划分标准上的识别结果为84.1%，在Cross-setup划分标准上的识别结果为86.0%，均保持在一个较好的水平，相比于其他方法呈现出了更为明显的优势，证明了本文提出的网络模型在性能上的优越性。

4　结　　论

如今基于视频的人体动作识别需求层出不穷。基于人体骨架数据的动作识别相比于RGB视频具有数据量小、无环境干扰等优点，从而成为了研究重点。本文从多特征、多分支以及多特征之间的关联为主线出发，首先基于输入的骨架数据，利用多种特征增强方法并通过改进的多分支融合特征输入到本文的神经网络中，然后使用多尺度时空卷积模块提取多尺度特征信息。在NTU RGB-D 60数据集的两种划分标准Cross-subject和Cross-view上的识别准确率分别为89.6%和95.1%，在NTU RGB-D 120数据集的两种划分标准Cross-subject和Cross-setup上的识别准确率分别为84.1%和86.0%，较好地提升了模型的动作识别准确率。

参　考　文　献：

[1] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]//2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Diego: IEEE, 2005: 886-893.

[2] DALAL N, TRIGGS B, SCHMID C. Human detection using oriented histograms of flow and appearance [C]//*9th European Conference on Computer Vision*. Graz: Springer, 2006: 428-441.

[3] WANG H, KLÄSER A, SCHMID C, *et al*. Action recognition by dense trajectories [C]//*Conference on Computer Vision and Pattern Recognition*. Colorado Springs: IEEE, 2011: 3169-3176.

[4] WANG H, SCHMID C. Action recognition with improved trajectories [C]//*Proceedings of the IEEE International Conference on Computer Vision*. Sydney: IEEE, 2013: 3551-3558.

[5] 钱慧芳，易剑平，付云虎. 基于深度学习的人体动作识别综述[J]. 计算机科学与探索，2021，15(3)：438-455.
QIAN H F, YI J P, FU Y H. Review of human action recognition based on deep learning [J]. *Journal of Frontiers of Computer Science and Technology*, 2021, 15(3): 438-455. (in Chinese)

[6] DU Y, FU Y, WANG L. Skeleton based action recognition with convolutional neural network [C]//2015 *3rd IAPR Asian Conference on Pattern Recognition*. Kuala Lumpur: IEEE, 2015: 579-583.

[7] WANG P C, LI W Q, LI C K, *et al*. Action recognition based on joint trajectory maps with convolutional neural networks [J]. *Knowledge-based Systems*, 2018, 158: 43-53.

[8] LI C, ZHONG Q Y, XIE D, *et al*. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation [C]//*Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Stockholm: ACM, 2018, 786-792.

[9] 李庆辉，李艾华，郑勇，等. 利用几何特征和时序注意递归网络的动作识别[J]. 光学 精密工程，2018，26(10)：2584-2591.
LI Q H, LI A H, ZHENG Y, *et al*. Action recognition using geometric features and recurrent temporal attention network [J]. *Optics and Precision Engineering*, 2018, 26(10): 2584-2591. (in Chinese)

[10] 李颀，邓耀辉，王娇. 基于轻量级图卷积网络的校园暴力行为识别[J]. 液晶与显示，2022，37(4)：530-538.
LI Q, DENG Y H, WANG J. Campus violence action recognition based on lightweight graph convolution network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2022, 37(4): 530-538. (in Chinese)

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

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

[12] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]//*Proceedings of the 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*. New Orleans: ACM, 2018: 7444-7452.

[13] LI M S, CHEN S H, CHEN X, *et al*. Actional-structural graph convolutional networks for skeleton-based action recognition [C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 3590-3598.

[14] YANG Z Y, LI Y C, YANG J C, *et al*. Action recognition with spatio-temporal visual attention on skeleton image sequences [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, 29(8): 2405-2415.

[15] SONG Y, ZHANG Z, SHAN C, *et al*. Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition [C]//*Proceedings of the 28th ACM International Conference on Multimedia*. Seattle: ACM, 2020: 1625-1633.

[16] 吴海滨,魏喜盈,刘美红,等. 结合空洞卷积和迁移学习改进YOLOv4的X光安检危险品检测[J]. 中国光学, 2021,14(6):1417-1425.
WU H B, WEI X Y, LIU M H, *et al*. Improved YOLOv4 for dangerous goods detection in X-ray inspection combined with Atrous convolution and transfer learning [J]. *Chinese Optics*, 2021, 14(6): 1417-1425. (in Chinese)

[17] 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*). Honolulu: IEEE, 2017: 1623-1631.

[18] LI S, LI W Q, COOK C, *et al*. Independently recurrent neural network (IndRNN): building a longer and deeper RNN [C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 5457-5466.

[19] LIANG D H, FAN G L, LIN G F, *et al*. Three-stream convolutional neural network with multi-task and ensemble learning for 3D action recognition [C]//2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*). Long Beach: IEEE, 2019: 934-940.

[20] ZHANG P F, LAN C L, ZENG W J, *et al*. Semantics-guided neural networks for efficient skeleton-based human action recognition [C]//2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle: IEEE, 2020: 1109-1118.

[21] LI S J, YI J H, FARHA Y A, *et al*. Pose refinement graph convolutional network for skeleton-based action recognition [J]. *IEEE Robotics and Automation Letters*, 2021, 6(2): 1028-1035.

[22] YOON Y, YU J, JEON M. Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition [J]. *Applied Intelligence*, 2022, 52(3): 2317-2331.

[23] CAETANO C, BRÉMOND F, SCHWARTZ W R. Skeleton image representation for 3D action recognition based on tree structure and reference joints [C]//2019 32*nd SIBGRAPI Conference on Graphics, Patterns and Images*. Rio De Janeiro: IEEE, 2019: 16-23.

[24] SONG Y F, ZHANG Z, WANG L. Richly activated graph convolutional network for action recognition with incomplete skeletons [C]//2019 *IEEE International Conference on Image Processing* (*ICIP*). Taipei, China: IEEE, 2019: 1-5.

[25] MEMMESHEIMER R, THEISEN N, PAULUS D. Gimme signals: discriminative signal encoding for multimodal activity recognition [C]//2020 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Las Vegas: IEEE, 2020: 10394-10401.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

作者简介：

张　磊（1996—），男，内蒙古呼和浩特人，硕士研究生，2019年于哈尔滨工程大学获得学士学位，主要从事计算机视觉、图像处理方面的研究。E-mail：zhangleiused@163.com

韩广良（1968—），男，山东嘉祥人，博士，研究员，2003年于中国科学院长春光学精密机械与物理研究所获得博士学位，主要从事图像和视频信息处理、目标识别与跟踪、机器视觉与人工智能等方面的研究。E-mail：hangl@ciomp.ac.cn

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net