基于双流骨架信息的人体动作识别方法

张　艳,肖文琛,张　博
(北华航天工业学院 计算机学院,河北 廊坊 065000)

摘　要:针对当前基于二维图像的人体动作识别算法鲁棒性差、识别率不高等问题,提出了一种融合卷积神经网络和图卷积神经网络的双流人体动作识别算法,从人体骨架信息提取动作的时间与空间特征进行人体动作识别。首先,构建人体骨架信息时空图,利用引入注意机制的图卷积网络提取骨架信息的时间和空间特征;其次,构建骨架信息运动图,将卷积神经网络网络提取到骨架运动信息的特征作为时空图卷积网络所提取特征的时间和空间特征的补充;最后,将双流网络进行融合,形成基于双流的、注意力机制的人体动作识别算法。算法增强了骨架信息的表征能力,有效提高了人体动作的识别精度,在 NTU-RGB+D60 数据集上取得了比较好的结果,Cross-Subject 和 Cross-View 的识别率分别为 86.5% 和 93.5%,相比其他同类算法有一定的提高。

关键词:动作识别;骨架信息;注意力机制;图卷积神经网络;双流网络
中图分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1673-629X(2024)01-0158-06
doi:10.3969/j.issn.1673-629X.2024.01.023

Human Action Recognition Method Based on Two-flow Skeleton Information

ZHANG Yan, XIAO Wen-chen, ZHANG Bo
(School of Computing, North China Institute of Aerospace Engineering, Langfang 065000, China)

Abstract: Aiming at the problems of poor robustness and low recognition rate of current human action recognition algorithms based on two-dimensional images, a two-stream human action recognition algorithm based on convolutional neural network and graph convolutional neural network was proposed to extract the temporal and spatial features of human action recognition from human skeleton information. Firstly, the spatial and temporal graph of skeleton information is constructed, and the graph convolution network with attention mechanism is used to extract the temporal and spatial characteristics of skeleton information. Secondly, the skeleton information action graph is constructed, and the features extracted from the convolutional neural network are used as the time and space features of the features extracted from the spatio-temporal graph convolutional network. Finally, the two-stream networks are fused to form a human action recognition algorithm based on dual flow and attention mechanism. The proposed algorithm enhances the representation ability of skeleton information and effectively improves the recognition accuracy of human movements. It achieves good results on the NTU-RGB+D60 data set, and the recognition rates of Cross-Subject and Cross-View are 86.5% and 93.5%, respectively, which is a certain improvement compared with other similar algorithms.

Key words: action recognition; skeleton information; attention mechanism; graph convolutional neural network; two-stream networks

0　引　言

人体动作识别是计算机视觉领域中一项复杂的任务,在智慧医疗、运动赛事、监控、人机交互等方面有着重要的社会应用价值[1]。人体动作识别根据数据源的不同可以分为基于 RGB 图像[2]的人体动作识别算法和基于骨架信息[3]的人体动作识别算法。基于 RGB 图像的人体动作识别算法鲁棒性差且易受外界环境的影响。相反,基于骨架信息的人体动作识别算法具有泛化能力强、不易受外界环境影响等优点。根据特征提取方式的不同,人体动作识别可以分为基于深度学习的人体动作识别算法和基于传统的机器学习的人体动作识别算法。基于深度学习的人体动作识别算法泛化能力相较于传统的机器学习的人体动作识别算法有了很大的提高。基于深度学习的人体动作识别算法可

收稿日期:2023-02-28　　**修回日期**:2023-06-29
基金项目:河北省高等学校科学技术研究项目(ZC2021006);廊坊市科技项目(2022011003);研究生创新基金项目(YKY202238)
作者简介:张　艳(1979-),女,副教授,硕导,博士,研究方向为图像处理、大数据与人工智能;通信作者:肖文琛(1997-),男,硕士研究生,研究方向为图像处理。

以分为基于 CNN(Convolutional Neural Networks)、基于 RNN(Recurrent Neural Network)的人体动作识别算法[4]。比如 Chen 等人[5]提出了一种基于序列的视点不变的方法来进行特征编码,并且将特征编码后的 RGB 图像通过多流 CNN 进行识别。文献[6]提出了基于注意力机制的 RNN 网络,分别赋予了不同帧之间骨骼节点不同的权重,进一步提高了动作的识别率。上述方法一定程度上提高了识别率,但是上述方法都是将骨架数据表示二维数据[7],不能完全表达骨骼关节点的时空信息。因此,从以上背景出发,该文提出了一种基于注意力机制的 AGCNS(Attention Graph Convolutional Networks)与 CNN 相结合的双流人体动作识别算法。主要创新点如下:提出了基于注意力机制的图卷积网络(AGCNS),通过 AGCNS 提取骨架信息的时空特征;结合骨架运动图进行时空特征的补充,提出了基于双流骨架信息的人体动作识别方法。

1　相关原理技术

1.1　骨架信息获取

随着微软 Kinect 度相机的问世和 OpenPose[8]算法的出现,提取人体动作的骨架信息不再是一件困难的事情。2012 年 2 月,微软正式发布了适合 Windows 平台的 Kinect 版本,并提供了 Kinect 开发包。通过 Kinect 开发包,配备 Kinect 相机可以提取到人体运动的三维骨架信息,这为基于三维骨架信息的动作识别提供了数据源,基于三维骨架信息的人体动作识别得到了进一步发展。OpenPose 人体姿态识别算法是一种有效检测图像中多人二维姿势的方法,是由美国卡耐基梅隆大学开发的开源库,可以实现实时的人体动作姿态的估计。该方法在首届 COCO2016 关键点挑战赛中排名第一,在性能和效率方面都大大超过了之前的最新结果,具有极好的鲁棒性。通过 OpenPose 进行姿态估计并提取骨架信息,为基于人体骨架信息动作识别奠定了基础。

1.2　卷积神经网络

神经网络是由大量人工神经元构成的、按照不同连接方式构建的网络。而卷积神经网络是神经网络中一种应用比较广泛的网络,主要应用在图像识别领域。卷积神经网络的结构主要可以分为三层:卷积层、池化层、全连接层。其中卷积层的作用是提取特征;池化层可以将无用的信息过滤掉,同时可以保留最显著的特征,这样大大减少了计算的复杂性;全连接层是一个完全连接的神经网络,主要作用是分类。

由何恺明等人提出的 ResNet[9],解决了深度网络退化问题,直到今天依旧有着广泛的应用场景。ResNet 的残差结构既不会增加参数,也不会增加模型复杂度。某种情况下,当上一层的输出结果达到最优时,在大多数情况下,恒等映射往往无法达到最优,这时就需要通过残差模块进行修正。ResNet 通过学习去拟合相对于上一层输出的残差,实验表明,ResNet 可以不断地增加网络的深度提高网络的性能,并且参数量更少,在众多数据集都有非常好的表现。

1.3　图卷积网络

传统的 CNN 在图像识别领域有较大的提升,CNN 的研究对象往往在有着规则空间结构的正方形栅格数据,比如图片数据,这些数据可以通过二维矩阵表示,很适合 CNN 进行处理。但是,现实生活中很多数据并不是有规则的空间结构,比如分子结构、脑神经结构以及人体骨骼点之间的连接关系。这些不规则的空间结构很难通过传统的 CNN 进行处理,这时可以通过图卷积网络来进行处理。图卷积的流程可以分为三步:第一步,将每个节点自身的特征信息经过转换发送给邻居节点;第二步,将每个图节点的邻居节点的信息进行聚合;第三步,将聚合后的信息做非线性变换,增加模型的表征能力。

图卷积网络的核心是图卷积操作,图卷积类似于 CNN 网络的卷积操作,作用是进行特征提取,具体公式为:

$$h_i^{l+1}=\sigma\left(\sum_{j\in N_i}\frac{1}{c_{ij}}h_j^l w_{R_j}^l\right) \tag{1}$$

其中,h_i^l 表示节点 i 在第 l 层的特征表达,c_{ij} 表示归一化因子,N_i 表示节点 i 的邻居,包含自身,R 表示节点 i 的类型,$w_{R_j}^l$ 表示 R_j 类型节点在第 l 层的变换权重参数。

2　模型结构

2.1　模型整体结构

双流网络是结合了图卷积神经网络和卷积神经网络,其中图卷积网络用来提取时间空间特征,卷积神经网络用来提取时空特征作为图卷积网络的补充。

提出的模型整体结构如图 1 所示。

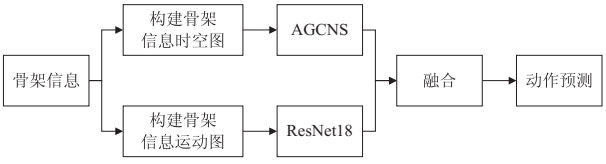

图 1　模型整体结构

首先,构建骨架信息时空图,然后将时空信息图送入基于注意机制的 AGCNS,得到动作预测结果。其次,将骨架信息特征编码为骨架信息运动图,将生成的 RGB 图像送入 ResNet18 中,得到动作预测结果。最后,将两个预测结果融合得到动作的预测类别。具体的,设 AGCNS 预测结果为 y_{AGCNS},ResNet 预测结果为 y_{ResNet},将双流的预测结果按权重进行相加,即 $y=$

$y_{\text{AGCNS}} + a * y_{\text{ResNet}}$，通过调参得到最好的识别效果。

2.2 构建骨架信息时空图

骨架序列是连续时间内二维骨架信息或三维骨架信息的集合。人体骨架可以看作为一个图的拓扑结构，在骨架信息的空间拓扑图结构加入时间信息就构成了时空信息图。本节遵循 Yan 等人[10]提出的动态骨架模型来构建时空骨架图，它可以通过自动学习骨架数据中的时空信息来克服以往方法的局限性。首先构建无向时空图，记为 $G=(V,E)$，其中 V 是节点集合，$V=\{v_{ti}\mid t=1,2,\cdots,T,i=1,2,\cdots,N\}$，$v_{ti}$ 表示一段时间内所有骨骼点的时空信息，其中 T 表示人体运动的总帧数，N 表示人体骨架中的所有关节点的数量；E 是边集，由 E_1 和 E_2 组成，其中 E_1 表示同一帧内骨骼点的连接，具体 $E_1=\{v_{ti}v_{tj}\mid (i,j)\in H\}$，其中 H 表示同一帧内所有骨架点的集合，E_2 表示不同帧之间同一骨架点之间的连接，$E_2=\{v_{ti}v_{(t+1)i}\}$。

根据已定义好的图 G，在空间维度上基于图形的卷积实现不像 2D 或者 3D 卷积那样简单，GCN 网络的图卷积网络定义的具体公式为：

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

其中，f_{in} 表示输入特征，其维度为 (C,V,T)，f_{out} 表示输出特征，$\boldsymbol{A}$ 表示单帧人体骨架的连接关系的邻接矩阵，单位矩阵 $\boldsymbol{I}$ 表示关节自连接，将多个输出通道的权重向量叠加，形成权重矩阵 $\boldsymbol{W}$，$\boldsymbol{\Lambda}$ 表示对角矩阵，具体公式为：

$$\Lambda^{ii}=\sum_j(A^{ij}+I^{ij}) \tag{3}$$

为了实现可学习边缘重要性加权，对于每个邻接矩阵，将其与一个可学习的权重矩阵 $\boldsymbol{M}$ 相伴。将等式中的矩阵 $(\boldsymbol{A}+\boldsymbol{I})*\boldsymbol{M}$，其中 $*$ 表示两个矩阵之间的元素乘积。上述公式可以被替换成：

$$f_{\text{out}}=\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})*\boldsymbol{M}\boldsymbol{\Lambda}^{-\frac{1}{2}}f_{\text{in}} \tag{4}$$

2.3 注意力机制

注意力机制是目前常用的数据处理方法，广泛用在图像识别、自然语言处理等不同的学习任务当中。人体动作识别过程中，不同帧之间的同一关节点的运动有一定的关联性，时间注意力机制，分别赋予不同帧之间不同关节点的时间权重，可以提高时空图卷积网络特征提取的能力。该时间注意力机制的结构如图 2 所示，具体公式为：

$$f_1=\sigma(M_t(\text{AvgPool}(f_{\text{in}}))) \tag{5}$$

其中，输入 f_{in} 特征为 $C\times T\times N$，AvgPool 表示平均池化，M_t 表示以一维卷积操作，σ 表示 Sigmoid 激活操作。一维卷积操作之后，通过一个 Sigmoid 函数获得 0～1之间归一化的权重得到 f_1，其特征大小为 $1\times T\times 1$，然后将 f_1 和 f_{in} 相乘并加入残差机制生成 f_{out}。

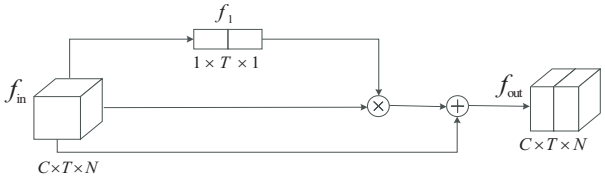

图 2 时间注意力机制

人体动作识别过程中，相同帧之间的同一关节点的运动有一定的关联性，同时不同帧的不同关节点之间也存在着一定关联性。空间变化程度的不同会影响动作识别的过程，因此本节引入了空间注意力机制。分别赋予不同关节点不同的权重，帮助时空图卷积网络更好地进行特征提取。该空间注意力机制结构和时间注意力机制结构类似，输入和输出特征相同。

通道注意力机制的目的是给不同的通道赋予不同的权重，增强模型的表征能力。具体结构如图 3 所示。

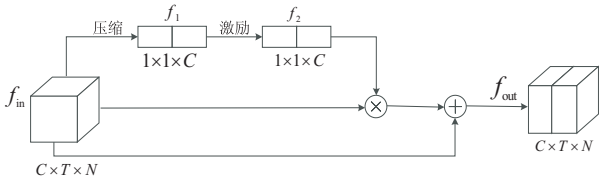

图 3 通道注意力机制

首先，将输入 f_{in} 特征（$C\times H\times W$）进行压缩操作，从空间维度来进行特征压缩，生成 f_1 将特征变成一个 $1\times 1\times C$ 的特征，得到的特征向量具有较强的全域性感受野，并且输出的通道数和输入的特征通道数相匹配，表示在特征通道上响应的全域性分布。具体公式为：

$$f_1=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{i=1}^{W}v_c(i,j) \tag{6}$$

其中，$v_c\in R^{H\times W}$，全局平均池化操作，从而使其具有全局的感受野，使得网络低层也能利用全局信息通过此步骤得到。其次是激励操作，通过加入参数 k_1 和 k_2 为每个特征通道赋予不同的权重，并通过 Sigmoid 进行归一化操作，得到 0 和 1 之间的权重 f_2，其特征大小为 $1\times 1\times C$，然后将 f_2 和 f_{in} 相乘并加入残差机制生成 f_1。具体公式为：

$$f_2=\sigma(\boldsymbol{k}_2\delta(\boldsymbol{k}_1f_1)) \tag{7}$$

其中，$\boldsymbol{k}_1\in R^{\frac{C}{r}\times C}$，$\boldsymbol{k}_2\in R^{C\times\frac{C}{r}}$ 分别为两个权重矩阵，δ 为 Relu 函数，σ 为 Sigmoid 激活函数。

2.4 基于注意力机制的 AGCNS

结合 2.2 节和 2.3 节，本节引入了注意力机制，提出的 AGCNS 模型基本单元的结构如图 4 所示。数据输入分别经过空间卷积层、归一化处理、激活处理、时间注意力层、空间注意力层、通道注意力层、时间卷积层，归一化处理、激活处理和失活处理得到输出特征，然后将原始输入特征和经过时空卷积后的输出特征相

加作为 AGCNS 模型单元的输出。空间卷积层的作用是提取空间特征信息,时间卷积层的作用是提取时间信息。其中,时间注意力层、空间注意力层以及通道注意力层顺序连接,并以残差结构的形式置于空间卷积层和时间卷积层的中间。

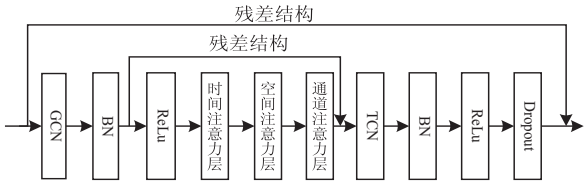

图 4　AGCNS 的基本单元

基于注意力机制 AGCNS 模型结构由上述 9 个基本单元构成,输入数据的通道为 3,前三个基本单元的输出通道为 64,步长为 1,中间三个基本单元的输出通道为 128,最后三个基于单元的输出通道为 256。经过 9 个基本单元后,将输出的特征图送入最大池化层和一个全连接层,最后经过 SoftMax 函数处理得到预测结果。

2.5　骨架信息特征编码为运动图

遵循文献[11]中的方法,该文将骨架信息转换为骨骼运动图骨架信息运动图,从而更好地提取时空信息特征。该运动图包含骨架的运动信息。首先,将深度第一树遍历顺序应用于骨架关节,以生成预定义的骨架链顺序 J,进而可以有效地保留原始骨架关节的空间信息。然后,将预定义的骨架链顺序 J 按照时间序列 T 的逐行堆叠得到矩阵 $\boldsymbol{M}$。其中矩阵 $\boldsymbol{M}$ 的大小为 $J \times T \times 3$, T 为骨架信息序列的总帧数,3 表示三维通道。根据矩阵 $\boldsymbol{M}$ 计算运动结构,具体公式为:

$$\boldsymbol{N}_{M,t} = \boldsymbol{M}_{J,t+d} - \boldsymbol{M}_{J,t} \tag{8}$$

其中,每个矩阵 $\boldsymbol{N}_{M,t}$ 由两个相差 d 帧的矩阵 $\boldsymbol{M}$ 计算差值而得,其大小为 $J \times (T-d) \times 3$。通过使用所提出的运动结构,建立了两种不同的表示:一种基于关节运动的大小,另一种基于关节运动的方向。使用以下公式计算两种表示:

$$A_{M,t} = \sqrt{(N^x_{M,t})^2 + (N^y_{M,t})^2 + (N^z_{M,t})^2} \tag{9}$$

$$S_{M,t} = \mathrm{stack}(S^{xy}_{M,t}, S^{yz}_{M,t}, S^{zx}_{M,t}) \tag{10}$$

$$S^{xy}_{M,t} = \tan^{-1}\left(\frac{N^y_{M,t}}{N^x_{M,t}}\right) \tag{11}$$

$$S^{yz}_{M,t} = \tan^{-1}\left(\frac{N^z_{M,t}}{N^y_{M,t}}\right) \tag{12}$$

$$S^{zx}_{M,t} = \tan^{-1}\left(\frac{N^x_{M,t}}{N^z_{M,t}}\right) \tag{13}$$

其中, $A_{M,t}$ 表示运动的大小, $S_{M,t}$ 表示三个运动方向的堆叠, $S^{xy}_{M,t}$ 表示坐标 x 和 y 的运动方向, $S^{yz}_{M,t}$ 表示坐标 y 和 z 的运动方向, $S^{zx}_{M,t}$ 表示坐标 z 和 x 的运动方向。运动大小 $\boldsymbol{A}$ 的维度为 $J \times (T-d) \times 1$;运动方向 $\boldsymbol{S}$ 的维度为 $J \times (T-d) \times 3$,3 为 (x,y,z) 三维方向。由于运动方向值会产生误差,可能没有运动也会产生运动的方向值,因此根据运动大小 $\boldsymbol{A}$ 对其进行过滤:

$$S'_{M,t} = \begin{cases} 0, & A_{M,t} < m \\ S_{M,t} & \text{otherwise} \end{cases} \tag{14}$$

其中, m 为运动大小的阈值, $S'_{M,t}$ 为过滤后的运动方向。

3　实验结果及分析

3.1　数据集

实验是在公共大型数据集 NTU RGB+D60[12] 上进行测试与验证。NTU RGB+D60 包含 60 个动作类型,共 56 880 个样本,其中有 40 类为日常行为动作,9 类为与健康相关的动作,11 类为双人动作。该数据集通过 3 台不同角度的 KinectV2 传感器采集获得,采集的数据形式包括深度信息、3D 骨骼信息、RGB 信息以及红外序列。这些样本都是由 40 名志愿者在特定的环境下进行采集的。其中每一帧骨架序列中的骨架序列包含 25 个关节,并且提供的注释给出了由 Kinect 深度传感器检测到的摄像机坐标系中的 3D 关节位置 (x,y,z)。数据集按照训练集和测试集划分的不同方式分为两类:

(1)交叉对象(Cross-Subject, CS):训练集包括 40 320个样本,测试集包括 16 560 个样本。其中,训练集来自同一个志愿者的动作,测试集来自剩余志愿者的动作。

(2)交叉视角(Cross-View, CV):训练集包括 37 920个样本,测试集包括 18 960 个样本。其中,训练集来自摄影机 2 号和 3 号,而测试集都来自摄影机 1 号。

遵循以上基准,验证所提算法的有效性。

3.2　实验细节

实验通过 Pytorch 深度学习框架进行验证,并在 Ubuntu18.04 系统,TeslaV100-PCIE、显存为 32 GB 的服务器上进行实验。将双流网络在一台服务器上分别进行训练,然后将预测结果融合得到最终的输出结果。

其中 ResNet18 包含 18 层,17 层卷积网络和 1 层全连接网络,在输出通道数为 64、步幅为 2 的 7×7 卷积层后,接步幅为 2 的 3×3 的最大池化层。ResNet18 使用 4 个由残差块组成的模块,每个模块使用若干个同样输出通道数的残差块。AGCNS 与 ResNet18 的初始学习率和 dropout 分别设置为 0.001 和 0.5,epoch 为 100,分类器都为 SoftMax,训练批次和测试批次设置为 64。

该文采用召回率和准确率作为模型评价的指标,召回率的具体公式为:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

准确率的具体公式为：

$$A = \frac{TP + TN}{TP + FN + FP + FN} \tag{16}$$

其中，TP，TN，FP，FN 分别代表真正例、真负例、假正例、假负例的样本个数。文中的混淆矩阵通过计算每个类别的召回率来衡量模型的效果，即混淆矩阵中对角元素表示预测值占真实值的百分比，通过混淆矩阵可以有效评估算法模型的视图变化和嘈杂等骨架挑战问题。

3.3 实验结果分析

首先，按照预设的参数训练 GCNS 网络。为了观察文中算法在 NTU RGBD+60 上的分类结果，采用混淆矩阵进行评估。混淆矩阵可以很清晰地观察文中模型在数据集各种不同动作的识别效果。作为后续实验结果的对比，且考虑到图像清晰的问题，给出初始化的 GCNS 的第 10 到第 30 类动作混淆矩阵，如图 5 所示。识别结果中，在 21 类动作中，有 11 类动作的识别率在 90% 及以上，8 类动作识别率大于 95% 甚至接近 100%，有 13 类动作识别率在 90% 以下，有 5 类动作识别率在 80% 以下，分别为 10-鼓掌，11-读书，12-写字，16-穿鞋，30-在键盘上打字。原因是极其相似的动作对识别率会产生一定的影响，比如读书和写字、在键盘上打字和玩手机，这些动作确实很难区分。

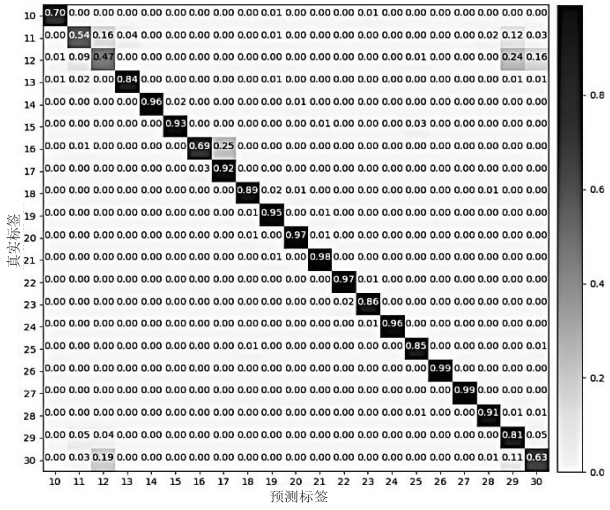

图 5 GCNS 的混淆矩阵

其次，为验证不同注意力机制对于 GCNS 的影响，进行了交叉实验，实验结果如表 1 所示。发现通过改进的基于注意力机制的 GCNS 一定程度上提高了识别率。相比于初始化 GCNS，加入通道注意力机制的 GCNS 在 Cross-Subject 和 Cross-View 情况下识别率分别提高了 2.5 百分点和 2.5 百分点，加入空间注意力机制的 GCNS 在 Cross-Subject 和 Cross-View 情况下识别率分别提高了 1.9 百分点和 1.8 百分点，加入时间注意力机制的 GCNS 在 Cross-Subject 和 Cross-View 情况下识别率分别提高了 2.0 百分点和 2.0 百分点，加入时间、空间、通道注意力机制的 GCNS 在 Cross-Subject 和 Cross-View 情况下识别率分别提高了 3.8 百分点和 3.1 百分点，表明加入以上三种注意力机制的 GCNS 可以有效提高动作识别率。该文将加入三种注意力机制的 GCNS 简称为 AGCNS。

表 1 注意力机制对 GCNS 网络的影响

算法	Cross-Subject /%	Cross-View /%
初始化 GCNS	81.6	88.7
加入通道注意力机制的 GCNS	84.1	91.2
加入空间注意力机制的 GCNS	83.5	90.5
加入时间注意力机制的 GCNS	83.6	90.7
加入时间、空间、通道注意力机制的 GCNS(AGCNS)	85.4	91.8

基于 AGCNS 和 ResNet18 网络中构成基于双流骨架信息的人体识别算法，如图 6 所示，给出第 10 到第 30 类动作的混淆矩阵。21 类动作中，有 16 类动作的识别率在 90% 以上，有 5 类动作识别率在 90% 以下。16 类动作当中有 9 类动作识别率大于 95% 甚至接近 100%，有 7 类动作识别率在 90% 到 95% 之间；5 类动作识别率在 90% 以下的分别为 10-鼓掌、11-读书、12-写字、29-玩手机、30-在键盘上打字；与初始化 GCNS 相比，识别率分别提高了 18 百分点，21 百分点，17 百分点，3 百分点，11 百分点。实验说明文中算法可以较好地区分相似动作且可以有效提高动作识别率。虽然有 5 类动作的识别率在 90% 以下，原因是极其相似的动作对识别率会产生一定的影响，但是文中算法依然有着很强的泛化能力，说明基于注意力机制的 AGCNS 和 ResNet18 双流人体动作识别方法在缺乏背景信息的情况下对于相似动作有着较好的识别效果。

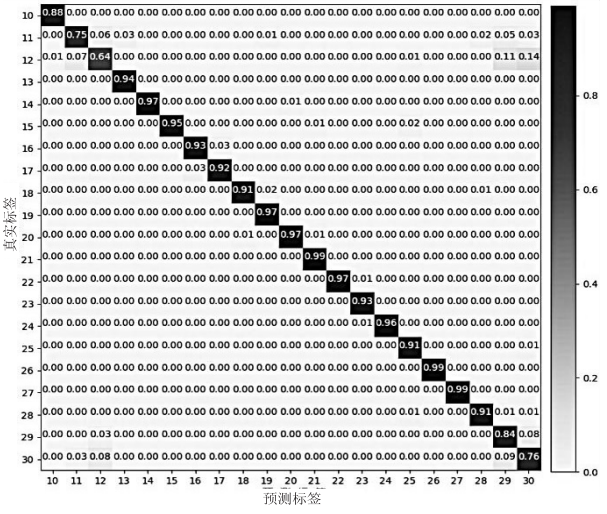

图 6 文中算法混淆矩阵

最后，为了验证算法的识别效果，将文中算法与国内外相关算法在 NTU RGB+D60 数据集上进行了对比。由于该数据集的约束性质，在训练文中模型时没

有任何数据增强。遵循训练集和测试集划分方式的不同分别进行不同的实验,分别验证 Cross-Subject 和 Cross-View 识别性能。对比结果如表 2 所示,相比其他算法,提出的双流动作识别方法在该数据集上效果更好,在 NTU RGBD+60 数据集上 Cross-Subject 和 Cross-View 的识别率分别达到了 86.5% 和 93.5%。

表 2　不同算法在 NTU RGBD+60 数据集上准确率的对比

算法	Cross-Subject/%	Cross-View/%
文献[10]	81.5	88.3
文献[12]	76.5	84.7
STA-LSTM[6]	73.4	81.2
VA-LSTM[13]	79.4	87.6
BGC-LSTM[13]	81.8	89.0
PR-GCN[14]	85.2	91.7
PeGCN[15]	85.6	93.4
文中算法	86.5	93.5

最后,为了更好地评价模型的训练速度,在特定的实验环境下,通过 100 个 epoch 的训练时间来衡量模型运算速度。如表 3 所示,文中算法的训练时间相比于 ST-GCN 的训练时间有了进一步的减少,说明文中算法在识别效率上有了一定的提升。

表 3　模型训练时间的比较

算法	训练时间/小时
ST-GCN	32
文中算法	30.5

4　结束语

针对当前基于二维图像的人体动作识别算法鲁棒性差、识别率不高等问题,引入了注意力机制和骨架信息运动图,提出一种基于 AGCNS 和 CNN 相结合的双流骨架信息人体动作识别方法。与传统基于 RGB 图像的人体动作识别方法不同,该文从人体骨架信息提取动作的时间与空间特征,利用加入注意机制的 AGCNS 网络提取骨架信息的时间和空间特征,同时通过 ResNet18 提取骨架信息运动图的时空特征,最后将两个网络进行融合,增强了骨架信息的表征能力,有效提高了人体动作的识别精度。该算法在 NTU-RGBD+60 数据集上取得了比较好的效果,Cross-Subject 和 Cross-View 的识别率分别为 86.5% 和 93.5%,相比其他同类算法,动作识别率有了一定的提高,同时模型训练也有一定的提升。

参考文献:

[1] 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.

[2] ZHU F,SHAO L,XIE J,et al. From handcrafted to learned representations for human action recognition:a survey[J]. Image and Vision Computing,2016,55:42-52.

[3] WEN Y H,GAO L,FU H,et al. Graph CNNs with motif and variable temporal block for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. Hawaii:AAAI,2019:8989-8996.

[4] 何秀玲,杨　凡,陈增照,等. 基于人体骨架和深度学习的学生课堂行为识别[J]. 现代教育技术,2020,30(11):105-112.

[5] LIU M,LIU H,CHEN C. Enhanced skeleton visualization for view invariant human action recognition[J]. Pattern Recognition,2017,68:346-362.

[6] SONG S,LAN C,XING J,et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]//Proceedings of the AAAI conference on artificial intelligence. San Francisco:AAAI,2017:31.

[7] BRUNA J,ZAREMBA W,SZLAM A,et al. Spectral networks and locally connected networks on graphs[J]. arXiv:1312.6203,2013.

[8] CAO Z,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. Hawaii:IEEE,2017:7291-7299.

[9] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:770-778.

[10] YAN S,XIONG Y,LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Thirty-second AAAI conference on artificial intelligence. New Orleans:AAAI,2018:5323-5332.

[11] CAETANO C,SENA J,BRÉMOND F,et al. Skelemotion:a new representation of skeleton joint sequences based on motion information for 3d action recognition[C]//2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS). Taipei,China:IEEE,2019:1-8.

[12] SHAHROUDY A,LIU J,NG T T,et al. Ntu rgb+ d:a large scale dataset for 3d human activity analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:1010-1019.

[13] ZHAO R,WANG K,SU H,et al. Bayesian graph convolution lstm for skeleton based action recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. Hawaii:IEEE,2019:6882-6892.

[14] LI S,YI J,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.

[15] 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.