文章编号：1005-5630(2022)04-0039-10 **DOI**: 10.3969/j.issn.1005-5630.2022.004.006

多尺度超图卷积骨架动作识别网络

秦晓飞[1]，赵 颖[1]，张逸杰[1]，杜睿杰[1]，钱汉文[1]，陈 萌[2]，张文奇[2]，张学典[1]
（1. 上海理工大学 光电信息与计算机工程学院，上海 200093；
2. 上海宇航系统工程研究所，上海 201109）

摘要：动作识别是计算机视觉基础任务之一，骨架序列包含了大部分的动作信息，因此基于骨架的动作识别算法受到很多学者关注。人体骨架在数学上是一个天然的图，所以图卷积被广泛应用于动作识别。但普通的图卷积只聚合两两节点间的低阶信息，不能建模多节点间的高阶复杂关系。针对此问题，本文提出一种多尺度超图卷积网络，在空间和时间两个维度聚合更丰富的信息，提高动作识别准确度。多尺度超图卷积网络采用编解码结构，编码器使用超图卷积模块聚合超边中多个节点间的相关信息，解码器使用超图融合模块恢复原始骨架结构，另外基于空洞卷积设计了多尺度时间图卷积模块以更好地聚合时间维度运动信息。NTU-RGB+D 和 Kinetics 数据集上的实验结果验证了算法的有效性。

关键词：动作识别；图卷积；超图卷积；空洞卷积
中图分类号：TN 391 **文献标志码**：A

Multiscale hypergraph convolutional network for skeleton-based action recognition

QIN Xiaofei[1]，ZHAO Ying[1]，ZHANG Yijie[1]，DU Ruijie[1]，
QIAN Hanwen[1]，CHEN Meng[2]，ZHANG Wenqi[2]，ZHANG Xuedian[1]
（1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Institute of Aerospace System Engineering of Shanghai, Shanghai 201109, China）

Abstract: Action recognition is one of the basic tasks of computer vision. The skeleton sequence contains most of the action information, so skeleton-based action recognition has attracted a lot of research attention. Mathematically, the human skeleton is a natural graph, so graph convolution is widely used in action recognition. But ordinary graph convolution only aggregates low-order information between pairwise nodes, and cannot model high-order complex relationships between multiple nodes. To solve this problem, a multiscale hypergraph convolutional network is proposed, which aggregates richer information in the two dimensions of space and time, so as to improve the accuracy of action recognition. The multiscale hypergraph convolutional network has an encoder-decoder structure. The encoder uses the hypergraph convolution module to aggregate relevant

收稿日期：2022-01-06
基金项目：上海市人工智能计划（2019-RGZN-01077）
作者简介：秦晓飞 (1982—)，男，高级工程师，研究方向为人工智能算法。E-mail：xiaofei.qin@foxmail.com

information between multiple nodes in the hyperedge, and the decoder uses the hypergraph fusion module to restore the original skeleton structure. In addition, a multiscale temporal graph convolution model based on dilated convolution is designed, which is used to better aggregate the temporal-dimension motion information. The experimental results on NTU-RGB+D and Kinetics datasets verify the effectiveness of this algorithm.

Keywords: action recognition；graph convolution；hypergraph convolution；dilated convolution

引　言

近年来，动作识别已成为计算机视觉领域的一个重要的分支，在人机交互、自动驾驶方面都有着广泛的应用。由于人类行为环境的复杂性，在执行动作识别任务时，经常受到相机移动、遮挡等复杂场景的干扰，限制了直接使用视频进行动作识别的方法的性能。随着深度相机的广泛应用和高性能姿态估计算法的出现，人们可以简单快速地获得人体骨架关节点位置信息。骨架关节点位置信息对于环境的干扰有较强的鲁棒性，因此基于骨架的动作识别算法取得了较好的效果，得到了动作识别领域越来越多的关注。

基于骨架的动作识别方法包括早期的手工特征设计方法[1-2]和近年来发展的基于深度学习的方法。手工特征设计方法由于其设计复杂、通用性差等原因，现在已基本不再使用。基于深度学习的骨架动作识别方法又分为卷积神经网络(convolutional neural network, CNN)类方法[3-5]和图卷积神经网路(graph convolutional network, GCN)类方法。CNN 类动作识别方法大多使用循环神经网络(recurrent neural network, RNN)[6-8]对骨架帧序列的时间和空间特征进行提取。虽然这类方法能够较好地描述时间维度特征，但对空间维度信息提取能力不足，主要原因是 CNN 类方法将骨架数据表示为向量序列或 2D 网格，不能完全表达关节之间的依赖性，忽略了人体的结构信息。数学上，人体骨架结构可以自然地看作以关节为顶点、以骨骼为边的图(Graph)，因此 GCN 可以有效地建模人体节点之间的结构信息，从而较好地提取人体的运动信息，虽然 GCN 直到近几年才被应用于骨架动作识别，但现已成为基于骨架动作识别任务的主流方法。

2018 年，ST-GCN[9]首次将 GCN 方法应用于骨架动作识别任务。它从时间和空间两个维度来处理骨架数据，较 CNN 类的方法取得了长足的性能提升，开创了基于 GCN 的骨架动作识别新领域。近三年的很多方法都是针对 ST-GCN 的改进[10-17]。ST-GCN 使用固定的邻接矩阵来表示人体的物理连接，对非物理连接节点间的互动信息提取能力不足。比如“拍手”这类动作，很大程度上依赖于左右手的互动，但骨架图上两手之间不存在直接的物理连接，ST-GCN 对此类动作识别效果较差。针对此问题，Dynamic GCN[12]提出了一种内容编码网络来自动地学习和更新节点间的连接关系；2s-AGCN[13]提出了一种自适应图卷积模块，该模块使用两个嵌入函数生成样本相关的关节点间连接程度 $\boldsymbol{C}$，并添加了一个可学习的邻接矩阵 $\boldsymbol{B}$，最后使用加法操作将原始邻接矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，$\boldsymbol{C}$ 相加得到一个自适应的邻接矩阵，取得了不错的效果。ST-GCN 只使用关节点坐标序列作为输入，信息来源较单一。针对此问题，ResGCN[11]和 2s-AGCN[13]分别提出了三流(节点流、骨骼流和速度流)和双流(节点流和骨骼流)输入的数据预处理方法，增加了模型信息来源，提高了动作识别准确度。

大多数现有的基于 GCN 的动作识别方法使用简单图描述人体连接关系，简单图的边只能连接两个节点，因此基于简单图的 GCN 层只能通过邻接矩阵学习节点间的低阶关系。然而，现实生活中人的动作往往需要多个节点相互配合才能完成，基于简单图的 GCN 网络需要堆叠多层才能描述这种高阶关系，但多层堆叠会导致过平滑、计算量大等问题。超图是简单图的扩展，超图的边可以连接多个节点，同一个节点可以属于不同的超边。因此将超图引入 GCN 动作识别网络可以较好地描述多节点间的关系。Hyper-GCN[18]首次尝试将超图网络引入骨架动作识别

领域，构造局部超边和全局超边提取高阶特征信息，并使用超图注意力机制获得相邻节点的不同权值。

受以上思想的启发，本文设计了一种用于骨架动作识别的多尺度超图卷积网络，主要贡献包括：首先将原始骨骼信息转换为节点序列、骨骼序列、动态序列分别输入多尺度超图卷积网络，形成一个三流网络，提高原始信息利用率；其次设计了一个以超图卷积模块为编码器、以超图融合模块为解码器的编解码结构,更好地建模多节点间的空间依赖关系；最后基于时间空洞卷积设计了一种多尺度时间图卷积模块，以建模动作的时间依赖关系。

1　算法

1.1　动作识别流程

动作识别的具体流程如图 1 所示。整个流程由输入数据预处理、多尺度超图卷积特征提取网络和预测分类三部分组成。对于输入的视频序列，人体关节点的三维坐标信息可由姿态估计算法得出。输入数据预处理部分，对人体关节点三维坐标 (x,y,z) 进行转换得到骨骼和动态数据。其中，骨骼可以表示为源关节点指向目标关节点的一个矢量，例如源关节点为 $v_1=(x_{v1},y_{v1},z_{v1})$、目标关节点为 $v_2=(x_{v2},y_{v2},z_{v2})$ 的骨骼可以表示为向量 $\boldsymbol{e}_{v1,v2}=(x_{v2}-x_{v1},y_{v2}-y_{v1},z_{v2}-z_{v1})$。动态数据表示连续帧之间的运动 $\boldsymbol{e}_{t1,t2}=(x_{t2}-x_{t1},y_{t2}-y_{t1},z_{t2}-z_{t1})$。将预处理后的关节坐标、骨骼和动态数据分别输入到三个独立训练的多尺度超图卷积网络中，每个流具有相同的网络结构。Softmax 分类器用来获得每个流的分类分数，最后将三个流的分类分数融合起来作为整个网络的预测结果。

1.2　多尺度超图卷积网络概述

本文提出的多尺度超图卷积网络结构如图 2 所示。该网络整体上属于一种三阶段的编解码结构 U 型网络，输入可以是关节、骨骼或动态数据。编码器部分使用两个本文设计的超图卷积模块（hypergraph convolution block, HCB）逐步减少特征维度，以聚集节点间的高阶信息；解码器部分使用两个本文设计的超图融合模块（hypergraph merging block, HMB）逐渐恢复原始骨架尺寸大小；编解码器之间采用跳级连接融合同阶段的编码器浅层信息与解码器深层信息。编码器和解码器的每个阶段都采用若干个自适应图卷积模块（adaptive graph convolution block, AGCB）来聚集同尺度特征的相邻节点信息。为了更好地建模输入序列帧间的相互依赖关系，设计了一种基于空洞卷积的多尺度时间图卷积模块（multiscale temporal graph convolution block, MTGCB）对解码器的输出特征进行处理。图 2 中模块下面的数

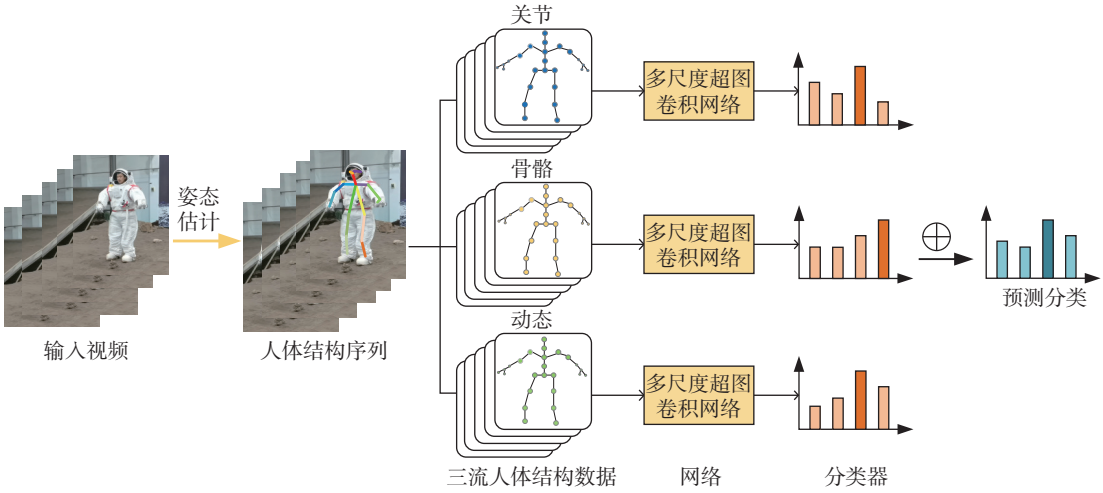

图 1　动作识别流程

Fig. 1　Action recognition process

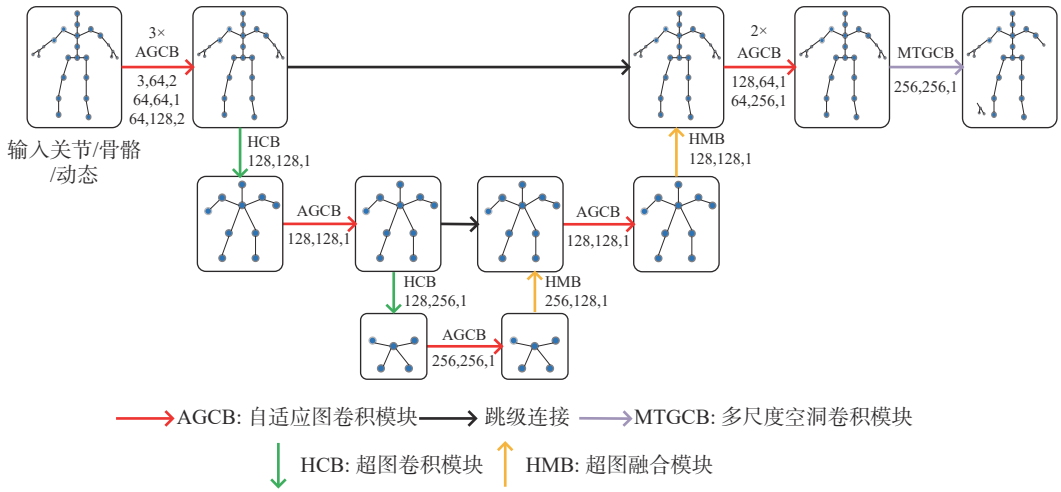

图 2　多尺度超图卷积网络结构

Fig. 2　Structure of multiscale hypergraph convolutional network

字三元组分别表示本模块的输入通道数、输出通道数、时间维度卷积步长。比如编码器第一阶段 AGCB 下面的(3,64,2)代表本 AGCB 的输入通道数是 3(即输入关节、骨骼或动态的三维数据)，输出通道数是 64，时间维度卷积步长为 2。

1.3　网络模块

1.3.1　自适应图卷积模块

多尺度超图卷积网络每个阶段的特征提取模块，本文借鉴了 2s-AGCN[13] 设计的 AGCB，AGCB 的结构如图 3 所示。在空间维度骨架数据具有不规则的空间结构，在时间维度骨架数据具有规则的几何结构，因此 AGCB 将骨架数据分为时间和空间两个维度进行特征提取。图 3 中的自适应图卷积网络(adaptive graph convolutional network, AGCN)用来聚集空间维度节点信息，时间卷积网络(temporal convolutional network, TCN)沿时间轴使用 3×1 卷积来聚集时间维度节点信息。这两个卷积层后面都有一个批归一化层(batch normalization, BN)和 Relu 激活层。此外为了增加 AGCB网络训练的稳定性，还使用了残差连接。

普通图卷积通常使用固定的物理连接关系来表示骨架，但是固定的物理连接缺乏对非相邻关节点依赖关系的建模能力，然而对于某些动作(比如拍手等)非相邻的关节点(左、右手等)间的依赖关系对动作的识别非常重要。针对此问题，图 3 中的 AGCN 部分通过卷积网络学习一个自适应邻接矩阵。不同于固定的物理连接，图的拓扑结构随着网络和参数一起优化，大大提高了模型的灵活性。依据输入数据的多样性，模型可以自适应地学习节点之间的拓扑结构。在动作

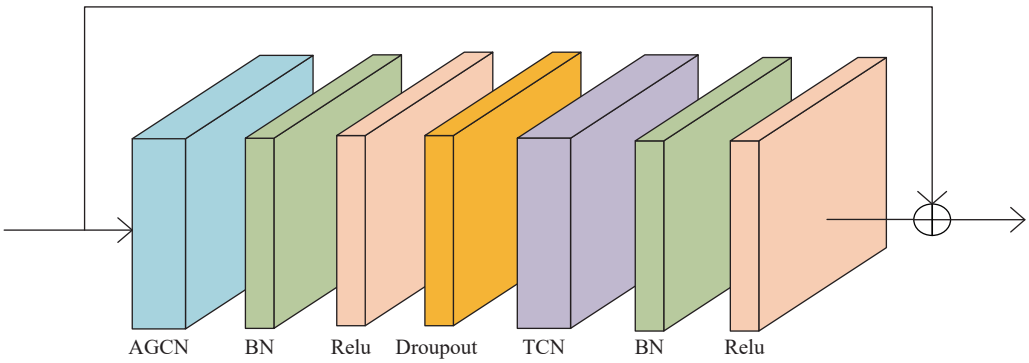

图 3　自适应图卷积模块

Fig. 3　Structure of adaptive graph convolution block

识别任务中，骨架被定义为图 $G=(V,E,\boldsymbol{A})$，其中 V 表示关节点的集合，E 表示边的集合，$\boldsymbol{A}\in\mathbf{R}^{N\times N}$ 表示骨架图的邻接矩阵，骨架图的特征由 (C,T,N) 的张量表示，其中 C 表示通道数，T 为时间长度，N 为关节点数量，则 AGCN 可表示为

$$f_{\mathrm{AGCN}}=\sum_{k}^{k_{\mathrm{v}}}\boldsymbol{W}_k(\boldsymbol{f}_{\mathrm{in}}(\boldsymbol{A}_k+\boldsymbol{B}_k+\boldsymbol{C}_k)) \tag{1}$$

式中：输入为 $\boldsymbol{f}_{\mathrm{in}}\in\mathbf{R}^{C_{\mathrm{in}}\times T\times N}$，输出为 $\boldsymbol{f}_{\mathrm{out}}\in\mathbf{R}^{C_{\mathrm{out}}\times T\times N}$；$\boldsymbol{A}_k$ 是原始的归一化邻接矩阵，$\boldsymbol{A}_k=\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Lambda}^{-\frac{1}{2}}$，其中，$\boldsymbol{\Lambda}_j^{ii}=\sum_k(A_j^{ik})+\alpha$，这里设置 $\alpha=0.001$ 来避免 $\boldsymbol{A}_j$ 空行；$\boldsymbol{B}_k$ 是一个参数化可学习的邻接矩阵，$\boldsymbol{C}_k$ 是样本相关的邻接矩阵，这几个邻接矩阵的维度都是 $\mathbf{R}^{N\times N}$；k_{v} 为分配策略数，AGCN 中将图的节点分为根节点、离心点和近心点三类，即 $k_{\mathrm{v}}=3$，k 为分配策略标号；$\boldsymbol{W}_k\in\mathbf{R}^{C_{\mathrm{out}}\times C_{\mathrm{in}}}$ 为可训练的权重矩阵。

1.3.2　超图卷积模块

人体动作是复杂多样的，像跳跃、站起、拍手等动作都需要多对关节点相互协调才能完成，因此建模多对关节点之间的高阶依赖关系对骨架动作识别任务至关重要。基于简单图的 GCN，无论其图结构是固定的还是自适应变化的，都很难描述这种多对关节点之间的高阶依赖关系。为此，本文将超图引入骨架动作识别任务，设计了一种编解码结构的多尺度超图卷积网络。编码器部分使用了两个超图卷积模块 HCB 来进行超边的融合，图 4 给出了本文设计的 HCB 在 NTU-RGB+D 和 Kinetics 两个数据集上的超边融合分配策略。由于超边可以包含多个关节点，超图卷积是对超边内多个关节点之间信息的聚合，因此 HCB 能够更好地建模多对关节点之间的依赖关系，加快关节点信息聚合的速度。HCB的计算过程如下。

首先定义超图的表示为 $G=(V,E,\boldsymbol{Q})$,其中 V 表示关节点的集合，E 表示超边的集合，$\boldsymbol{Q}$ 表示超图卷积的关联矩阵，$\boldsymbol{Q}\in\mathbf{R}^{N\times M}$。本文解码器中两个 HCB 中用到的 $\boldsymbol{Q}$ 可分别根据图 5 所示的两层超边融合分配策略得到，当超边 ε_j 连接节点 v_i 时，则 $Q_{ij}=1$，否则 $Q_{ij}=0$。超图卷积利用关联矩阵来聚集超边内多个关节点间的信息。

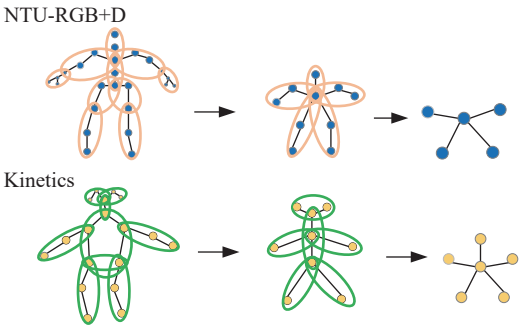

图 4　超边融合的分配策略

Fig. 4　Allocation strategy for hyperedge merging

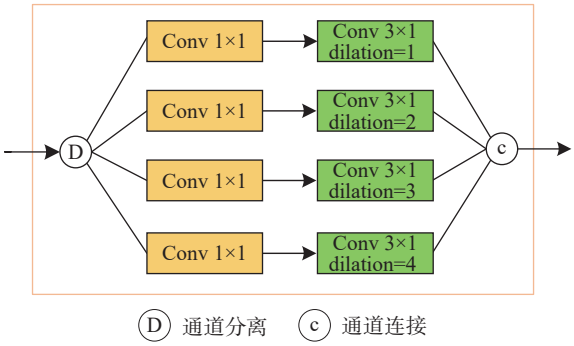

图 5　多尺度时间图卷积模块

Fig. 5　Structure of multiscale temporal graph convolution block

为了防止超边多次融合后信息爆炸，本文使用标准化超图连接，即通过归一化使节点的最大连接度不大于 1，对于 N 个节点和 M 个超边的超图，其标准化超图连接度的计算方法如下：

$$\boldsymbol{H}=\boldsymbol{D}_v^{-\frac{1}{2}}\boldsymbol{Q}\boldsymbol{W}_\varepsilon\boldsymbol{D}_\varepsilon^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}_v^{\frac{1}{2}} \tag{2}$$

式中：$\boldsymbol{D}_v\in\mathbf{R}^{N\times N}$ 是对角化超图节点度矩阵，其对角元素表示该节点连接超边的个数；$\boldsymbol{D}_\varepsilon\in\mathbf{R}^{M\times M}$ 是对角化超图超边度矩阵，其对角元素表示该超边内节点的个数；$\boldsymbol{W}_\varepsilon$ 表示超图超边之间的权重矩阵。类似图卷积定义的方式，本文利用标准化超图连接 $\boldsymbol{H}$ 与超图关联矩阵 $\boldsymbol{Q}$ 的矩阵乘积作超图卷积操作，可得 HCB 的计算公式如下：

$$\boldsymbol{f}_{\mathrm{HCB}}=\sigma(\boldsymbol{W}_1\boldsymbol{f}_{\mathrm{in}}\boldsymbol{H}\boldsymbol{Q}) \tag{3}$$

式中：$\boldsymbol{f}_{\mathrm{HCB}}\in\mathbf{R}^{C_2\times T\times M}$ 是 HCB 的输出特征；$\boldsymbol{f}_{\mathrm{in}}\in\mathbf{R}^{C_1\times T\times N}$ 是 HCB 的输入特征；σ 是一个非线性激活函数；$\boldsymbol{W}_1\in\mathbf{R}^{C_2\times C_1}$ 是一个可以训练的参数矩阵。

1.3.3　超图融合模块

HCB 使空间维度的特征图变小、感受野增大，解码器部分需要恢复特征的空间分辨率。图像领域通常用反卷积和反池化等上采样方法获取更高分辨率的特征图，然而这些方法并不适用于没有规则空间结构的图网络。为此，本文基于 HCB 的一种逆运算，设计了一种超图融合模块 HMB。HMB 的主要作用有两点：(1)编码器部分进行 HCB 操作后，图的空间维度变小，这意味着如果不进行上采样操作，同阶段解码器部分的图的空间维度将无法与编码器特征对齐，从而无法通过跳级连接进行特征融合。所以 HMB 的第一个作用是使编解码结构同阶段的空间特征图的维度对齐；(2)HMB 可以学到人体不同部分(即不同超边)的重要性，例如拍手动作，人的手这部分的重要性比较高，HMB 可通过权重参数对人的手所涉及的关节点进行加权增强。

类似图卷积定义的方式，本文利用标准化超图连接 $\boldsymbol{H}$ 与超图关联矩阵 $\boldsymbol{Q}^{\mathrm{T}}$ 的矩阵乘积作超图卷积操作，可得 HMB 的计算公式如下：

$$\boldsymbol{f}_{\mathrm{HMB}}=\sigma(\boldsymbol{W}_2\boldsymbol{f}_{\mathrm{in}}\boldsymbol{H}\boldsymbol{Q}^{\mathrm{T}}) \tag{4}$$

式中：$\boldsymbol{f}_{\mathrm{HMB}}\in\mathbf{R}^{C_4\times T\times N}$ 是 HMB 的输出特征；$\boldsymbol{f}_{\mathrm{in}}\in\mathbf{R}^{C_3\times T\times M}$ 是 HMB 的输入特征；σ 是一个非线性激活函数；$\boldsymbol{W}_2\in\mathbf{R}^{C_4\times C_3}$ 是一个可以训练的参数矩阵。

对于编解码结构的同一阶段，编码器部分输出的特征包含丰富的细节信息，解码器部分输出的特征包含丰富的高阶信息，融合两部分的特征可为后续动作识别分类提供更丰富的信息。为此，本文采用跳级连接和逐元素相加对编解码器的特征进行融合。

$$\boldsymbol{f}_{\mathrm{out}}=\boldsymbol{f}_{\mathrm{HMB}}+\boldsymbol{f}_{\mathrm{AGCB}} \tag{5}$$

式中：$\boldsymbol{f}_{\mathrm{out}}$ 为融合后的特征；$\boldsymbol{f}_{\mathrm{HMB}}$ 为 HMB 的输出特征；$\boldsymbol{f}_{\mathrm{AGCB}}$ 为同阶段编码器自适应图卷积模块的输出特征。

1.3.4　多尺度空洞图卷积模块

HCB 和 HMB 在空间维度获得了更大的感受野，但缺乏对时间维度信息的描述。虽然 AGCB 中的 TCN 操作使用了 3×1 卷积来聚集时间维度节点信息，但本文提出的多尺度超图卷积网络层数较少，其中仅包含 8 个 AGCB，在时间维度上的建模能力是有限的。有些方法[19]为了获得时间维度上较大的感受野将卷积核扩大，但这样会导致计算量大大增加。针对此问题，本文在 AGCB 的基础上，设计了一种多尺度时间图卷积模块 MTGCB，其结构是使用图 5所示的通道分离多尺度空洞卷积模块代替图 3 所示 AGCB 中的 TCN 模块。

MTGCB 首先使用 AGCN 对输入特征的空间维度信息进行聚合，之后将 AGCN 输出的特征按通道维度平均分成 4 份，即图 5 中所示的通道分离操作，这样可以减少模块的计算量。然后不同分支采用1×1卷积进行通道信息融合，使用空洞率分别为 1、2、3、4 的3×1空洞卷积获得不同时间跨度的运动信息。最后将不同分支提取的特征级联起来给最后的动作分类网络使用。

2　实验

本部分在 NTU-RGB+D[20] 和 Kinetics[21] 两个大规模动作识别数据集上验证本文提出的多尺度超图卷积网络(multiscale hypergraph convolutional Network, MHCN)。

2.1　数据集

NTU-RGB+D[20] 是一个著名且广泛使用的动作识别数据集，由 56 880 个动作剪辑、60 个动作类和 4 000 000 帧组成，包括日常动作、互动动作和与健康有关的动作。他们邀请了 40 名志愿者进行数据收集工作。3 个相同高度不同水平视角的深度摄像机同时捕捉同一动作，3 个深度摄像机的水平视角分别为 45°、0°、−45°。数据集包含每个志愿者 25 个关节点的 3D 位置。每个视频中最多包含 2 个人。NTU-RGB+D 数据集通常使用 CS 精度(Cross Subject Accuracy)和 CV 精度(Cross View Accuracy)来评价模型性能。

Kinetics[21] 是一个大规模且重要的人体动作识别数据集，包括 30 万个 YouTube 视频剪辑，共有 40 个动作种类。视频剪辑分为训练集(240 000

个剪辑）和验证集（20 000 个剪辑）。数据集使用 OpenPose[22] 姿态估计算法得到人体骨架序列，每个人有 18 个关节点，每个关节点由其在像素坐标中的二维坐标 (x,y) 及其置信度得分 s 组成，最终表示为 (x,y,s)。Kinetics 数据集通常使用 TOP1 和 TOP5 精度来评价模型性能。

2.2　实验细节

模型是使用 PyTorch 框架搭建的，使用交叉熵作为损失函数，优化方法采用带惯量的梯度下降，惯量系数为 0.9，权重衰减系数 0.000 1，批量大小为 64。对于 NTU-RGB+D 数据集，每个序列最多包含 2 人，当人数不足 2 人时，使用 0 填充操作将输入数据扩充为 2 人。另外该数据集的每个序列最多包含 300 帧，当帧数少于 300 帧时，使用重复填充将其扩充为 300 帧。初始学习率设置为 0.1，在第 30 个 epoch 和第 40 个 epoch 时下降至 0.01，共训练 60 个 epoch。对于 Kinetics 数据集，每个序列包含 150 帧，每帧中包含 2 个人体骨架。初始学习率设置为 0.1，在第 45 个 epoch 和第 55 个 epoch 时下降至 0.01，总训练次数同样为 60 个 epoch。

2.3　消融分析

为了验证本文所提出的各模块的有效性，在 NTU-RGB+D 数据集上进行消融分析。首先验证本文所提出的 HCB 和 HMB 的有效性，为了进行公平的比较，本文在 2s-AGCN 基础上，通过修改输入为三流，修改 2s-AGCN 最后一个 AGCB 为 MTGCB，得到基准算法。然后在基准算法基础上逐渐添加 10 节点的 HCB、HMB 和 5 节点的 HCB、HMB。实验结果如表 1 所示。表 1 中 $+\varepsilon_{10}$ 代表在 Baseline 的第 3 个 AGCB 之后添加一个 HCB，在第 6 个 AGCB 之后添加一个 HMB，并使用跳级连接进行特征融合；$+\varepsilon_5$ 代表在 Baseline 的第 4 个 AGCB 之后添加一个 HCB，在第 5 个 AGCB 之后添加一个 HMB，并使用跳级连接进行特征融合。从表 1 结果可知，添加 HCB 和 HMB 后，网络性能有所提升，说明 HCB 和 HMB 能够有效地融合超边内的多对关节点之间的信息。

表 1　HCB 和 HMB 的消融分析
Tab. 1　Ablation study of HCB and HMB

方法	CV精度/%
Baseline	95.3
Baseline+ε_{10}	96.0
Baseline+ε_{10}+ε_5	97.1

为了验证不同骨架输入数据对结果的影响，本文使用所设计的多尺度超图卷积网络分别进行了多种单流、两流、三流对比实验，实验结果如表 2 所示。表 2 中的 w/o 表示“没有”的意思，比如 w/o 骨架表示三流中除去骨架流，只剩下关节和动态两流输入。从表 2 可以看出，两流的方法比单流方法效果好，三流方法比两流方法效果好，这表明每个输入数据分支对模型性能提高都是必要的。从“w/o 动态”两流方法的结果可知，去掉动态输入流后模型精度降低了 1.9%，性能下降非常明显，这表明本文添加的动态输入流数据中包含了很多具有动作分辨力的信息。

表 2　不同骨架输入数据对结果的影响
Tab. 2　Comparison of results obtained via different skeleton input data

结构	数据	CV精度/%
三流	骨架、关节、动态	97.1
两流	w/o骨架	96.0
	w/o关节	95.8
	w/o动态	95.2
单流	骨架	93.7
	关节	93.5
	动态	92.1

为了验证 MTGCB 中不同空洞率的效果，本文进行了不同空洞率组合的实验，表 3 列出了实验结果。如表 3 所示，当 4 个分支的时间空洞率都设置为 1 时，MTGCB 就退化成了 AGCB；增大 4 个分支的时间空洞率可以增大时间维度的感受野，从而提高模型的表现，但当空洞率大于 3 时，模型表现开始下降，这说明不同时间空洞率都能够提取一定的动作信息。本文所提方法在 MTGCB 4 个分支上分别使用不同时间空洞

率，并将 4 个分支的结果通过级联融合，从而可以提取多种时间尺度上的动作信息，如表 3所示，达到了最优的效果。

表 3　不同空洞率下模型的表现
Tab. 3　The performance of models with different dilation factors

空洞率	CV精度/%
1	95.1
2	95.6
3	96.1
4	95.9
1, 2, 3, 4	97.1

图 6 所示为本文算法在 NTU-RGB+D 数据集上的学习曲线，其中左 y 轴表示的是训练精度，右 y 轴表示的是训练损失。由图 6 可知在训练过程中，随着 epoch 的增加，模型的训练精度逐渐提高，训练的损失则逐渐减少。

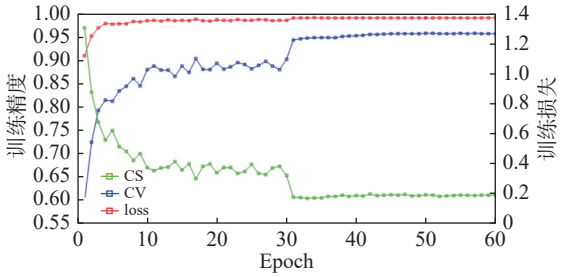

图 6　多尺度超图卷积网络在 NTU-RGB+D 数据集上的学习曲线

Fig. 6　Learning curve of multiscale hypergraph convolutional network on NTU-RGB+D dataset

2.4　对比实验

为了验证所提方法的优越性，将多尺度超图卷积网络 MHCN 和当前主流的骨架动作识别方法在 NTU-RGB+D 和 Kinetics 数据集上进行比较。表 4 和表 5 分别给出了各模型在 NTU-RGB+D 和 Kinetics 数据集上的表现。相较于当前最优模型，MHCN 在 NTU-RGB+D 数据集上，CS 精度提高了 1.1%，CV 精度提高了 0.9%；MHCN 在 Kinetics 数据集上，TOP1 精度提高了 1%，TOP5 精度提高了 1.7%。

表 4　在 NTU-RGB+D 数据集上与最新方法的比较
Tab. 4　Comparison with state-of-the-art methods on the NTU-RGB+D dataset

方法	CS精度/%	CV精度/%
Lie-Group[2]	50.1	52.8
TCN[23]	74.3	83.1
ST-GCN[9]	86.8	94.2
AS-GCN[14]	86.8	94.2
2s-AGCN[13]	88.5	95.1
SGN[17]	89.0	94.5
AGC-LSTM[7]	89.2	95.0
DGNN[24]	89.9	96.1
Res-GCN[11]	90.0	96.0
SGCN[16]	90.1	96.2
MHCN	91.2	97.1

表 5　在 Kinetics 数据集上与最新方法的比较
Tab. 5　Comparison with state-of-the-art methods on the Kinetics dataset

方法	TOP1精度/%	TOP5精度/%
TCN[23]	20.3	40.0
ST-GCN[9]	30.7	52.8
AS-GCN[14]	34.8	56.5
DGNN[24]	36.9	59.6
2s-AGCN[13]	36.1	58.7
Hyper-GCN[18]	37.1	60.0
SGCN[16]	37.1	60.1
MHCN	38.1	61.8

3　结论

骨架动作识别任务中，简单图不能很好地建模多个关节点之间的高阶信息，为此本文将超图引入骨架动作识别任务，设计了以超图卷积模块为超边融合算法、以超图融合模块为骨架尺寸恢复算法的编解码结构多尺度超图卷积骨架识别网络。该网络同时将关节、骨骼、动态三流数据作为输入以充分利用输入信息。该网络中的多尺度时间图卷积模块，使用不同的时间空洞率提取不

同时间跨度的动作信息。消融分析验证了本文所提各模块的有效性，对比实验验证了本文所提方法的优越性。

参考文献：

[1] GOWAYYED M A, TORKI M, HUSSEIN M E, et al. Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition[C]//Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing: IJCAI, 2013.

[2] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 588 – 595.

[3] DING Z W, WANG P C, OGUNBONA P O, et al. Investigation of different skeleton features for CNN-based 3D action recognition[C]//Proceedings of 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Hong Kong, China: IEEE, 2017: 617 – 622.

[4] LI C, ZHONG Q Y, XIE D, et al. Skeleton-based action recognition with convolutional neural networks[C]//Proceedings of 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Hong Kong, China: IEEE, 2017: 597 – 600.

[5] LI C K, WANG P C, WANG S, et al. Skeleton-based action recognition using LSTM and CNN[C]//Proceedings of 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Hong Kong, China: IEEE, 2017: 585 – 590.

[6] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016: 816 – 833.

[7] LIU J, WANG G, DUAN L Y, et al. Skeleton-based human action recognition with global context-aware attention LSTM networks[J]. IEEE Transactions on Image Processing, 2018, 27(4): 1586 – 1599.

[8] SI C Y, CHEN W T, WANG W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1227 – 1236.

[9] 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: AAAI, 2018.

[10] CHEN Y X, MA G Q, YUAN C F, et al. Graph convolutional network with structure pooling and joint-wise channel attention for action recognition[J]. Pattern Recognition, 2020, 103: 107321.

[11] SONG Y F, ZHANG Z, SHAN C F, 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. Virtual Event: ACM, 2020: 1625 – 1633.

[12] YE F F, PU S L, ZHONG Q Y, et al. Dynamic GCN: context-enriched topology learning for skeleton-based action recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia. Virtual Event: ACM, 2020: 55 – 63.

[13] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12026 – 12035.

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

[15] LI M S, CHEN S H, ZHAO Y H, et al. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 211 – 220.

[16] YANG W J, ZHANG J L, CAI J J, et al. Shallow graph convolutional network for skeleton-based action recognition[J]. Sensors, 2021, 21(2): 452.

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

[18] HAO X K, LI J, GUO Y C, et al. Hypergraph neural network for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 2263 – 2275.

[19] LIU Z Y, ZHANG H W, CHEN Z H, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 140 – 149.

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

[21] KAY W, CARREIRA J, SIMONYAN K, et al. The Kinetics human action video dataset[Z]. arXiv: 1705.06950, 2017.

[22] 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(1): 172 – 186.

[23] KIM T S, REITER A. Interpretable 3D human action analysis with temporal convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 1623 – 1631.

[24] SHI L, ZHANG Y F, CHENG J, et al. Skeleton-based action recognition with directed graph neural networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7904 – 7913.

（编辑：张　磊）