控制与决策
Control and Decision

基于时域扩张残差网络和双分支结构的人体行为识别

薛盼盼, 刘云, 李辉, 陶冶, 田嘉意

引用本文:
薛盼盼,刘云,李辉,陶冶,田嘉意. 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.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0648

基于时域扩张残差网络和双分支结构的人体行为识别

薛盼盼[1], 刘 云[1], 李 辉[1,2†], 陶 冶[1], 田嘉意[1]

(1. 青岛科技大学 信息科学技术学院，山东 青岛 266061；
2. 智能感知与自主控制教育部工程研究中心，北京 100124)

摘 要: 图卷积网络由于能够直接处理关节点拓扑图在行为识别方面表现出较好的性能而备受关注,但是这类方法中经常存在长时信息依赖建模能力较弱以及未关注空间语义与时间事件变化不均衡问题,对此,提出基于时域扩张残差网络和双分支结构的人体行为识别方法. 在时空行为特征提取方法中,不仅用图卷积提取空间域特征,而且用扩张因果卷积和残差连接结构来构建时域扩张残差网络以提取时域特征,该网络能够在未大量增加参数的基础上有效扩大在时域上的感受野,从而更好地获得在时域上的人体关节信息的长时依赖关系. 同时构建双分支结构,其中低帧率分支以较少的时间帧数和较多的通道数侧重于提取丰富的空间语义信息,高帧率分支以较多的时间帧数和较少的通道数在保证网络轻量级的前提下有效捕捉人体行为的快速变化. 实验结果表明,所提出方法在NTU RGB+D数据集上的准确率高于目前先进的行为识别方法.

关键词: 图卷积；行为识别；扩张卷积；残差连接；双分支结构

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0648 **开放科学(资源服务)标识码(OSID):**

引用格式: 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.

Human behavior recognition based on time domain extended residual network and dual branching structure

XUE Pan-pan[1], *LIU Yun*[1], *LI Hui*[1,2†], *TAO Ye*[1], *TIAN Jia-yi* [1]

(1. College of Information Science and Technology，Qingdao University of Science and Technology，Qingdao 266061，China；2. Engineering Research Center of Intelligence Perception and Autonomous Control of Ministry of Education，Beijing 100124，China)

Abstract: The graph convolution network has attracted much attention because it can directly process the topological graph of joint points and has good performance in behavior recognition. However, this kind of methods often have the problems of weak long-term information dependence modeling ability and not paying attention to the imbalance between spatial semantics and temporal events. Therefore, a human behavior recognition method based on the time-domain extended residual network and a dual-branch structure is proposed. In the method of spatiotemporal behavior feature extraction, not only the graph convolution is used to extract spatial domain features, but also the extended causal convolution and the residual connection structure are used to construct the time-domain extended residual network to extract time-domain features. The network can effectively expand the receptive field in time domain without increasing a large number of parameters, so as to better obtain the long-term dependence of human joint information in time domain. At the same time, a dual branch structure is constructed, in which the low frame rate branch focuses on extracting rich spatial semantic information with less time frames and more channels, while the high frame rate branch focuses on capturing the rapid changes of human behavior with more time frames and less channels. The accuracy on the NTU RGB + D data set is higher than the current advanced behavior recognition methods.

Keywords: graph convolution；behavior recognition；dilated causal convolution；residual connection；double branch structure

收稿日期: 2021-04-16；**录用日期:** 2021-08-18.

基金项目: 智能感知与自主控制教育部工程研究中心开放基金项目(K100052021006)；国家自然科学基金项目(61702295)；山东省高等学校优秀青年创新团队计划项目(2019KJN047).

责任编委: 张文安.

†通讯作者. E-mail: lihui@qust.edu.cn.

0 引 言

近年来,行为识别因其在视频理解中的重要作用而被广泛研究. 在深度学习中,可以利用多种数据进行行为特征的提取,比如: RGB数据、深度数据、关节点数据等. 在以往的研究中大多是利用前两种数据,如文献[1-3];关节点数据由于其噪声较小,更有利于提高行为识别的准确率[4-7]. 人体关节可以表示为拓扑图,图卷积网络能够处理这种拓扑图并取得了较好的效果. 但是在这些研究中,多数作者更关注于空间行为特征的提取,而忽略了长时信息依赖的重要性.

在静态图像$p(x,y)$的识别中,一般是对x和y两个空间维度做对称性的处理,这是因为根据对静态图像的统计表明,图像近似具有各向同性和位移不变,所谓的各向同性就是所有的方向都是相同的. 但是,视频关节点数据可以表示为$p(m(x,y,z),t)$. 其中: $m(x,y,z)$表示关节点的位置坐标,t表示所处的时间帧. 视频数据在方位x、y、z上是对等的,但是,时间和空间方向是不对等的,这体现在对于视频行为识别中空间的语义信息往往是变化缓慢的. 因此,在行为识别中关注空间语义与时间事件变化的不均衡性对于提高准确率是非常重要的.

针对以上问题,本文提出时域扩张残差网络和双分支特征提取方法,并且在此基础上设计基于时域扩张残差网络和双分支结构的人体行为识别方法. 通过实验对比分析,表明了本文方法的有效性.

1 国内外研究现状

目前,常见的处理关节点数据的深度学习的方法有卷积神经网络(convolutional neural network, CNN)和图卷积网络(graph convolutional network, GCN),对应的关节点数据的表示方式为伪图像和拓扑图.

基于CNN的方法分别将时间帧和骨架关节的位置坐标编码为行和列,然后将数据馈送到CNN中进行行为识别,类似于图像分类. Li等[8]将关节点数据的视频信息标准化到0～255的范围,从而映射到彩色图像中,然后采用多尺度深卷积神经网络(CNN)结构和微调策略来提高性能. Zhang等[9]提出了一种视图自适应方案,该方法能够有效节省用于设计复杂的关节点数据预处理标准以处理各种情况的人力资源. Caetano等[10-11]、Li等[12]从设计新的骨架表示图入手. 其中: Caetano等在文献[10]中提出了一种基于树结构和参考关节的3维行为识别的骨架表示方法;在文献[11]中又引入了一种新的方法,通过计算骨架关节的运动幅度和方向值来编码时间动态,使用不同的时间尺度来计算关节的运动值,能够有效过滤噪声运动值; Li等[12]采用集合代数的方式对骨架关节信息进行重新编码. Yang等[13]提出了一个由多个卷积神经网络组合而成的轻量级网络框架,大大提高了速度. 以上研究虽然有效挖掘了关节点的运动特征,但是都难以表达不同关节之间的依赖关系.

为了更加有效地表示关节之间的依赖关系,人们开始关注关节点的图形结构. 随着图卷积网络在计算机视觉领域的推广,人们开始将该方法应用于行为识别. Yan等[14]首先应用图卷积来处理关节点数据:在空间域上,他们根据人体关节在物理骨架上进行自然连接;在时域上,将连续帧中的相同关节进行连接. 同时根据图结构的特殊性,提出了一种新的构造图卷积层的方法,这种方法是利用基于距离的采样函数确定每个关节点的邻居集合,在邻居集中进行图卷积操作,然后将其用作构建最终时空图卷积的基本模块. 但是, Yan等的工作未考虑某些关节物理上无直接连接而在行为分类上具有相关性的情况. Li等[15]提出的时空图卷积方法也存在相似的缺点. Li等[16]提出了一种编码器-解码器的方法来捕获隐含的关节相关性,以及使用邻接矩阵的高阶多项式获取关节之间的物理链接. Gao等[17]将图形回归用于基于骨架的行为识别,该方法能够优化时空帧的基础图形,充分利用人体关节之间空间上物理和非物理的依赖关系以及连续帧上的时间连通性. Tang等[18]提出了深度渐近强化学习方法,该方法可以提取关键帧,然后用图卷积网络进行行为识别,行为识别的准确率较高. Li等[19]不仅利用关节相关性获得更丰富的关节依赖关系,而且扩展了现有的关节拓扑图来表示更高阶的依赖关系,同时增加了动作预测的功能. Shi等[20]提出了一种双流自适应图卷积网络,将人体的关节点信息作为第1阶信息,将人体的骨骼信息作为第2阶信息,同时在这两种信息上进行建模. 在所构建的自适应网络中,通过在空间上参数化另外两种类型的图形表明了在人体物理骨架上无自然连接的关节之间可能存在的数据相关性.

以上这些利用图卷积进行行为识别的研究中均存在长时信息依赖建模能力较弱的问题,并且多数网络框架是在空间与时间特征上提取串联的模式,在该模式中默认对于所有的动作空间和时间维度完全同等重要,并且只能处理一种帧率,但是,空间语义与时间事件变化通常具有不均衡性. 针对以上问题,本文所做的主要贡献如下:

1) 利用扩张卷积和残差连接结构来构建时域扩张残差网络,该网络能够扩大在时域上的感受野,从

而更好地获得人体关节点信息的长时依赖关系,提高行为识别的准确率.

2) 构建双分支结构,该结构中低帧率分支以较少的时间帧数和较多的通道数侧重于提取丰富的空间语义信息,高帧率分支以较多的时间帧数和较少的通道数保证网络轻量级的前提下有效捕捉人体行为的快速变化,同时使用将高帧率分支中的特征向低帧率分支中融合的方法使网络能够端到端地训练,其中融合采用混合组卷积的方式,可以减少计算量并且促进两个分支的特征之间的信息流动.

2　基于时域扩张残差网络和双分支结构的人体行为识别方法

本文提出的人体行为识别方法如图 1 所示. 主要分为两个阶段: 首先,视频以关节点数据的形式分别输入进基于关节点数据的行为识别网络和基于骨骼数据的行为识别网络;然后,分别计算分类得分再做加权双流特征决策,进而得出最终行为识别的结果.

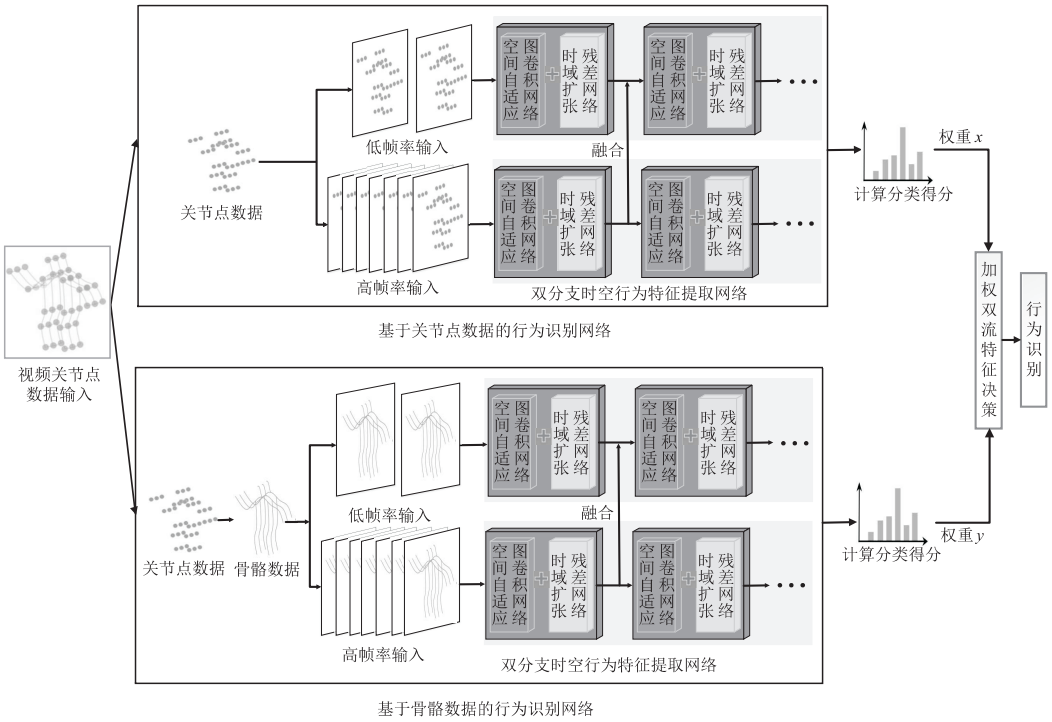

图 1　本文提出的人体行为识别方法技术路线图

图 1 中的双分支时空特征提取网络由低帧率分支和高帧率分支组成,每个分支均由 10 个时空行为特征提取基本模块组成,每个模块主要包括空间自适应图卷积网络和时域扩张残差网络,同时还包括 Dropout 层、Relu 激活层、BN (BatchNorm) 标准化层. 两个分支所对应的基本模块进行特征融合,融合采用混合组卷积的方式,可以减少计算量并促进两个分支的特征之间的信息流动.

2.1　时空特征提取基本模块

2.1.1　空间自适应图卷积网络

本文使用时空图来模拟关节沿空间域和时域的拓扑信息[14]. 每个关节点的属性特征是空间坐标向量 (x, y, z),使用每一个人关节 3D 坐标来表示骨架序列. 对于单个人的单帧骨架可以表示为 $G = (V, E)$. 其中: V 是关节点的集合, E 是边的集合. 针对一系列帧的骨架可以表示为 $F = G_1, G_2, \ldots, G_T$,其中 T 表示视频总共的帧数. 空间自适应图卷积如下所示:

$$H_{\text{out}} = W H_{\text{in}} M. \tag{1}$$

其中: W 表示图卷积的权重矩阵; H_{in} 和 H_{out} 分别表示输入和输出的特征图矩阵; M 表示注意力矩阵,注意力矩阵能够为不同边指定不同的重要性等级. 在式 (1) 中令 $M = A + P$, A 表示原始邻接矩阵,若两个关节在人体中相连,则对应的元素是 1,否则为 0. 矩阵 P 中的元素大小与矩阵 A 相同,初始化为 0,在训练过程中可以更新. 这种方法不仅能够改变现有边的重要性,还可以添加新的边,因为某些特定的动作中无直接相连的关节也有可能具有相关性. 这里 A 相当于用一个固定的拓扑图作为人体先验知识对模型进行正则化处理,避免了只有 P 的模型因过于灵活而容易收敛到局部最优的情况.

2.1.2 时域扩张残差网络

为了提取时域的行为特征信息,使用时域扩张残差连接的方式进行操作,每次完成一个关节点,完成一个关节点后进行下一个关节点的特征提取. 本文所提出的时域扩张残差网络能够有效地扩大在时域上的感受野,即在未大幅度增加参数量的同时能够提取到更多的人体行为在时域上的特征.

抓取动作的长时依赖信息对人体行为识别非常重要,正如RNN模型的循环自回归结构,它能够对时间序列进行很好的表示. 但是,RNN模型在进行训练时需要较多的内存,而且不同时间段上共用参数容易导致梯度消失和梯度爆炸的问题. 同时,由于RNN前一个时刻隐藏层的状态会参与到后一个时刻的计算过程而无法并行计算,难以缩短模型训练时长. 与传统的RNN模型相比,时域扩张残差网络能通过扩大感受野的方式抓取长时的依赖信息,同时可以避免RNN模型所要面临的问题.

时域扩张残差网络算法流程如下:

1) 输入:空间域行为特征图

2) for $(n=0;n<N;n++)$ // n表示第n层,N表示总层数

3) for $(i=0;i<2;i++)$

4) Relu; // 激活层

5) WeigtNorm; // 参数规范化

6) BatchNorm; // 归一化

7) Dilated Causal Conv; // 扩张因果卷积

8) if $(n>0$ & channel$(n)!=$channel$(n-1))$ // channel(n)表示第n层中的通道数

9) 1∗1 Conv // 调整通道数

10) 输出:时域行为特征图.

图2是时域扩张残差网络框架,该网络能够在时域上提取行为特征. 为了像RNN一样,能够将数据映射到相同长度的输出序列,该网络采用了添加零填充的方式以达到保证隐藏层与输入层长度相等的效果. 同时为了实现增大感受野的效果,采用了扩张因果卷积. 残差连接结构在两层之间代替层与层之间的简单连接,这样能够防止网络较深时容易出现网络退化现象. 由于两层之间的通道数可能不一样,这里设计了1∗1卷积来对通道数进行调整. 图2左边部分$h_{ij}=(x_{ij},y_{ij},z_{ij})$表示扩张因果卷积输入的第$i$个关节点在第$j$个时刻的特征向量,$s_{ij}=(x'_{ij},y'_{ij},z'_{ij})$表示扩张因果卷积输出的第$i$个关节点第$j$个时刻的特征向量. 扩张等价于在每个卷积核中的每两个相邻的元素之间引入一个固定的步长d代表扩张系数,当$d=1$时,扩张因果卷积减小为传统的规则卷积[21]. 使用更大的扩张系数可以使顶层的输出代表更大范围的输出,从而有效地扩大卷积操作的感受野,同时卷积核的实际大小并未改变,所以参数量不会大幅度增加.

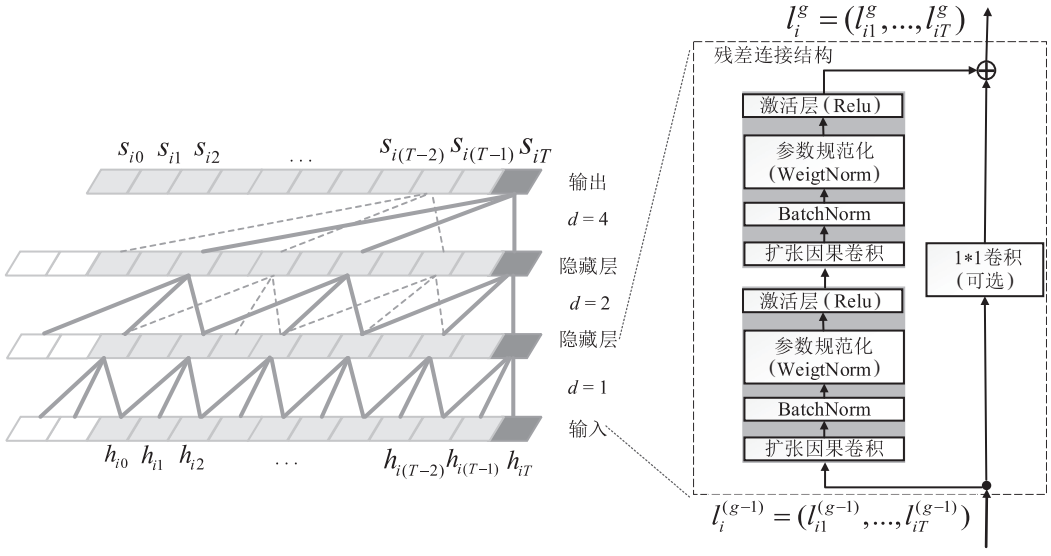

图 2 时域扩张残差网络框架

下面是扩张因果卷积的定义:卷积核函数为$F=(f_1,f_2,\dots,f_K)$,K表示卷积核的大小. 序列$H=(h_1,h_2,...,h_T)$,在h_t处,扩张系数为d的扩张因果卷积可以表示为$(F_{*d}H)_{h_t}=\sum_{k=1}^{K}f_k h_{t-(K-k)d}$. 时域扩张残差网络中的扩张卷积的感受野大小为$R_n\times 1$,其中$R_n$表示第$n$个隐藏层的感受野,$R_n=R_{n-1}+(K-1)\times d_n$[22],$d_n$表示第$n$层的扩张系数. 这说明感受野的大小取决于扩张卷积层数和卷积核大小. 扩张系数会随着网络层数呈指数形式增加,如图2中的

扩张系数为1、2、4,在实际应用中可以根据实验结果确定扩张网络层数.

图2的右边是残差连接结构[23]的细节图,其中 $l_{ij}^{g} = (x_{ij}^{g}, y_{ij}^{g}, z_{ij}^{g})$ 表示第 i 个关节第 j 帧时刻的关节向量信息. x_{ij}, y_{ij}, z_{ij} 分别表示第 i 个关节第 j 帧时刻的三维坐标点的位置. 其中使用BatchNorm可以降低数据之间的绝对差异,通过归一化对数据做去相关性. 同时BatchNorm能够使参数量不至于过大,故可以作为一种正则化的方式代替其他正则化方法,如Dropout等. 但是,当batchsize的值较小时,BatchNorm的效率会降低,因此,再加入参数规范化函数WeightNorm以得到更好的效果. 与BatchNorm在数据的层面上做归一化不同,参数规范化是在权值维度上做归一化,并且与样本量无关,没有额外的参数,也会更节约显存.

由于时域扩张残差网络的感受野取决于扩张卷积层数和卷积核大小,更深的网络的稳定性变得更加重要. 但是,简单地增加深度容易导致梯度爆炸或梯度消失,从而产生不稳定性. 为了解决这一问题,通常采用将权重参数初始化和中间层正则化的方法,但是,随着层数的增加还会出现网络退化的问题. 为了解决网络退化,在时域扩张残差网络中加入残差连接结构[24],因为解决网络退化问题可以等价于解决如何让网络的冗余层产生恒等映射. 图2中右边的连线即为残差连接,当维度不匹配时,通过1∗1的卷积操作使得维度达到匹配状态.

时域扩张残差网络具有以下优点:

1) 在时域上提取关节特征时具有并行性,输入序列可以在时域卷积模块中作为一个整体来处理.

2) 该网络可以灵活地扩大在时域上的感受野,感受野的大小由扩张卷积层数和卷积核大小所决定.

3) 该网络与传统的卷积神经网络一样,不存在不同时间段上共享参数的问题,因而不易发生梯度爆炸或梯度消失.

4) 时域扩张残差网络中的卷积核是共享的,能够降低内存的使用量.

2.2　双分支特征提取

先前的工作是将原始的视频关节点数据未经处理直接输入到行为识别网络,这时不会遗漏掉任何时间帧上的信息,同时为了尽可能多地提取行为特征,每个帧上要有足够的特征提取通道,这样会使计算量增大,尤其是当使用以图卷积为基础的特征提取方法时,因为图卷积本身计算复杂度较高.

为了缓解高帧率和高通道数与计算量之间的矛盾,将原本的时空特征提取网络设计成双分支的. 其中:一个分支输入的是低帧率的视频数据,在低帧率分支上尽可能多地提取空间语义信息,设置更多的通道数;另外一个分支输入的是高帧率的视频数据,在这个分支上虽然输入的数据量多于低帧率分支,但是,该分支更加侧重于获取高时间分辨率下的快速变化的运动,所以设置的通道数较少,因此计算量明显小于低帧率分支的. 通过这样两个不同帧率分支的设计能够更加有效地探索不同时间速度的潜力.

首先介绍低帧率分支,该分支所输入的视频数据是在原始视频的基础上按照一定的时间步长进行采样得到的. 如果每 γ 帧中取一帧,原始视频的帧数是 T,则经过采样处理后输入进低帧率分支中的视频帧数可以表示为 T/γ. 与低帧率分支相对应的是高帧率分支,为了使高帧率分支具有时间保真度,高帧率分支使用原始数据作为数据. 当高帧率分支与低帧率分支的帧数之比为 a 时,低帧率分支在采样视频帧时,每 a 帧取一帧,这里的 $a > 1$. 最终低帧率分支中视频输入的帧数可以表示为 T/a.

如果高帧率分支的通道数表示为 C_i,则低帧率分支的通道数可以表示为 bC_i,其中 C_i 表示第 i 个时空特征提取基本模块的通道数. 低帧率分支中每个时空特征提取基本模块的通道数如表1所示. 高帧率分支时空特征基本模块的通道数是低帧率分支通道数的 $1/b$ 倍,b 的具体取值情况由实验决定.

表 1　低帧率分支时空特征提取基本模块通道数

时空特征提取基本模块层数	输入通道数	输出通道数
第1层	3	64
第2层	64	64
第3层	64	64
第4层	64	64
第5层	64	128
第6层	128	128
第7层	128	128
第8层	128	256
第9层	256	256
第10层	256	256

通道数低可以理解为在空间维度的建模能力较弱,减弱低帧率分支中的空间建模能力,增强高帧率分支中的时间建模能力也是在综合考虑计算量以及识别准确率之后所选择的一种折中方案.

为了使训练在双分支的网络中能够端到端地进行,需要在每个时空特征提取基本模块之间采用横向连接的方式融合两个分支的特征,从而使分支可以得到另一个分支上学习到的表征. 当两个分支进行融合时,如果每个分支的时间帧数(即时间域维度)相

等,则融合很容易做到. 但是很明显,高帧率分支与低帧率分支中时间帧数是不同的,所以需要将其时间域维度调整至相同,然后再进行融合.

为了更好地表示两个分支在融合前数据维度的变换情况,将数据表示为{时间帧数,通道数}的格式. 高帧率分支中数据维度可以表示为$\{aT/\gamma, C_i\}$,低帧率分支中数据维度可以表示为$\{T/\gamma, bC_i\}$.

本文采用将高帧率分支特征压入低帧率分支横向连接融合的方式,如图3所示. 高帧率分支采用时间步长卷积的方式调整数据维度变为$\{T/a, C_i\}$. 调整好数据,便可将高帧率分支提取到的特征送入低帧率分支. 在进行融合时有两个输入和一个输出,两个输入分别是低帧率分支中未经调整时间维度的数据和高帧率分支中已经调整时间维度的数据,输出直接作为低帧率分支的下一个时空特征提取基本模块的输入.

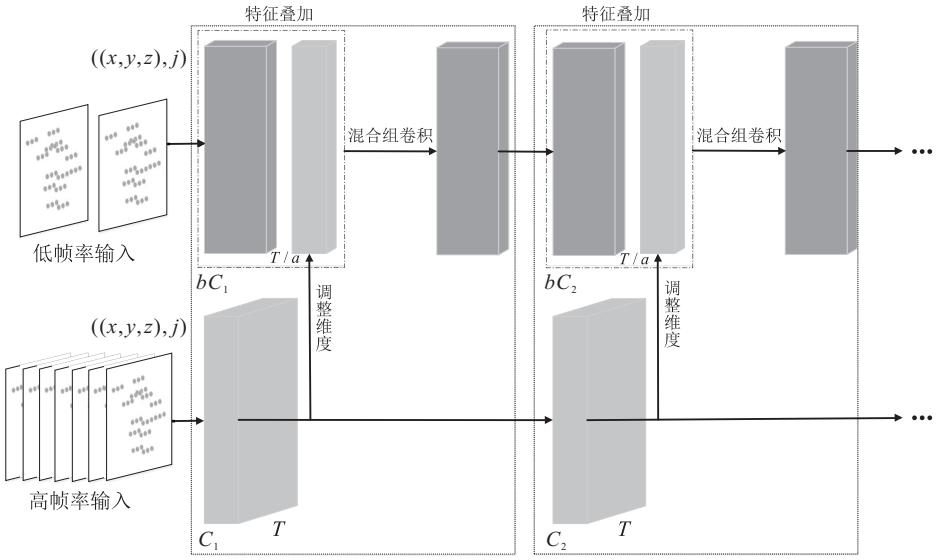

图3 横向连接融合示意图

在融合阶段每一个时间帧都是对应的,用x_t^h表示高帧率分支的第t帧的输入,用x_t^l表示低帧率分支第t帧的输入,于是连接融合可以表示为

$$y_t = \mathrm{cat}(x_t^h, x_t^l). \tag{2}$$

即在对应的时间帧上将两个分支中的特征通道进行串联叠加,这样便可保证信息的完整性. 如果这时直接将得到的数据输入进下一个时空特征提取基本模块,则通道数相当于两个分支通道数之和,10次融合会使参数量明显增加. 因此,希望采用一定的方法使融合后的通道数与融合前低帧率分支中的通道数相等,也为C_i.

一般类似于双分支结构,在融合调整通道数时使用1∗1的卷积,假设希望得到的通道数是C,则需要使用的卷积核的个数也是C. 但是,当通道数很多时,用这种方式会大幅度增加参数量,所以将采用分组卷积的方式降低计算量,如图4所示.

在图4中:左边(浅色部分)表示低帧率分支中的特征通道,其通道数为bC_i;右边(深色部分)表示高帧率分支中的特征通道,其通道数为C_i. 假设将低帧率分支的bC_i个通道分成N组,将高帧率分支的C_i个通

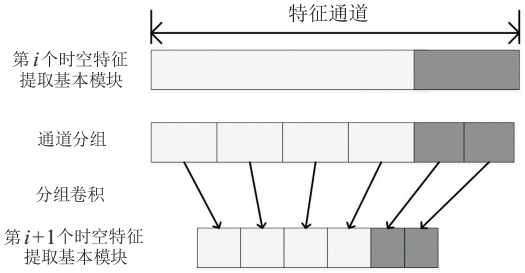

图4 分组卷积示意图

道分为N/b组,每个组内进行k个1∗1的卷积,则最终卷积得到的通道数为$(N + N/b)\times k$. 当按照实际情况将k的值进行适当调整后即可降低特征通道数至bC_i. 这样1∗1的卷积就可以在小组内的通道内进行而不是所有的通道内进行,从而降低计算量.

由于将通道分组,在进行1∗1的卷积中只能将所在小组内的通道进行融合,从而会造成不在同一个小组内的信息无法流通,这不利于最终的行为识别结果. 为了解决这个问题,应用通道混合的思想,也就是将原先所划分的小组再适当划分成更小的子组,子组之间相互混合,然后再进行分组卷积,如图5所示,这样每个组之间的信息能够更好地融合高帧率分支与低帧率分支中的特征.

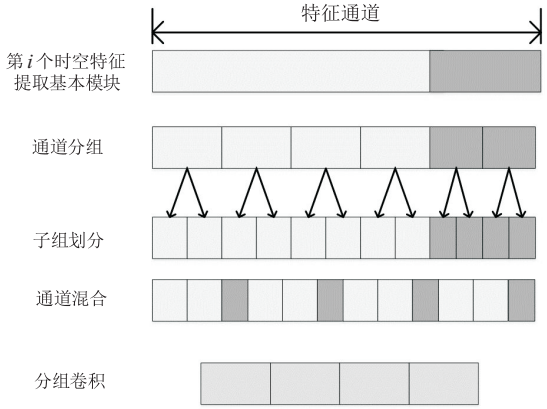

图5　混合分组卷积示意图

2.3　双流特征决策

基于关节点数据的行为识别网络与基于骨骼数据的行为识别网络中的时空行为特征提取网络是相同的,但是,由于一般关节点数据集中仅仅包含原始的关节点信息,基于骨骼信息的行为识别网络需要增加一个对关节点信息进行处理而得到骨骼数据的环节. 骨骼定义为源关节指向目标关节的向量,包含骨骼的长度信息和方向信息,该信息对于行为识别也很重要. 源关节为靠近骨骼重心的关节,目标关节为远离重心的关节[20]. 例如,源关节为$v_1=(x_1,y_1,z_1)$,目标关节为$v_2=(x_2,y_2,z_2)$,骨骼向量的计算结果为$e_{v1,v2}=(x_2-x_1,y_2-y_1,z_2-z_1)$.

图1中的加权设计需要提前训练好两个基于不同数据流的网络;然后,分别将关节点数据和骨骼数据输入时空行为特征提取网络,用softmax函数[25]计算出两个行为识别网络的分类得分;最后,将得分加权相加,得到融合分数并预测行为标签[26]. 两个基于不同数据流的行为识别网络可提供识别结果互补,提高行为识别的准确率. 在进行分数相加融合时,由于两个网络所用数据不同而导致性能不同,在进行相加融合时的权重也是不同的,具体不同网络的权重设置根据实验确定.

3　实验结果及分析

3.1　数据集简介

为了验证本文所提出的行为识别方法的性能,在NTU RGB+D这一个常用的标准数据集上进行测试. 该数据集用3个Microsoft Kinect v2传感器采集,这3个传感器分别放置的角度是−45°、0°、45°,一共包含60类动作. 其中第1～49个动作是单人动作,第50～60个动作是双人交互动作. 每个动作都由多人进行表演,并且从3个角度进行拍摄. 在NTU RGB+D数据集上进行训练和测试一般有两种模式:一种是跨表演者模式,表示为CV;一种是跨视角模式,表示为CS. 本文所涉及的行为识别的准确率均为Top1的准确率. 在跨表演者模式中,一半样本用于训练,一半样本用于测试;在跨视角模式中,将两个视角的样本用于训练,一个视角的样本用于测试.

3.2　参数设置及网络训练

实验硬件环境为GPU 1080Ti(4核),显存为256 G,CUDA版本为8.0. 所用开源框架为PyTorch 1.0,Python版本为3.6. 为了克服过拟合的问题,在训练时采用权重衰减方法,权重衰减率设置为0.000 1. 迭代次数设置为90,学习率设置为0.01,其中,当迭代到70次时学习率除以10.

3.3　实验结果分析

时域扩张残差网络中扩张卷积的层数会影响感受野的大小,本文将时域扩张残差网络中扩张卷积的卷积核设置为:当扩张卷积层数为1、2、3、4、5时,扩张系数依次为1、2、4、8、16,随着扩张系数的增大感受野也随之增大. 表2为不同的扩张卷积层数所对应的实验结果. 从表2中可以看出,随着网络层数的加深,感受野的增大,行为识别的准确率也会增加,其中当扩张卷积层数为5时,行为识别准确率最高. 但是,为了平衡计算量与有益影响,网络不能无限制地加深. 下文所有实验都将扩张卷积的层数设置为5. 表2中在调整不同扩张卷积层数时,两个网络的权重与表3中实验所得到的最优的权重设置一致.

表2　不同扩张卷积层数所对应的实验结果

扩张卷积层数	CS/%	CV/%
1	82.12	91.04
2	85.61	93.08
3	86.18	93.82
4	87.90	94.75
5	**88.71**	**95.62**

表3　基于不同数据的网络权重设置的实验结果

基于关节点数据的网络权重	基于骨骼数据的网络权重	CS/%	CV/%
4.0	6.0	88.54	95.39
4.2	5.8	88.60	95.42
4.4	**5.6**	**88.71**	95.44
4.6	5.4	88.64	94.47
4.8	5.2	88.60	95.52
5.0	5.0	88.57	95.53
5.2	**4.8**	88.44	**95.62**
5.4	4.6	88.50	95.54
5.6	4.4	88.44	95.53
5.8	4.2	88.36	95.52
6.0	4.0	88.25	95.49

本文提出的人体行为识别方法是由基于关节点数据的人体行为识别网络和基于骨骼数据的人体行为识别网络分别训练然后加权融合得到的. 在CS模

式和CV模式中两个网络的权重设置是不同的,表3是加权双流特征决策两个行为识别网络权重设置的实验结果. 可以看出:在CS模式中,基于关节点信息的人体行为识别网络和基于骨骼数据的人体行为识别网络的权重分别设置为4.4和5.6时所得到的实验结果最好;在CV模式中,分别设置为5.2和4.8时所得到的实验结果最好. 本文中所有涉及加权双流决策的权重的设置均与表3中的最优设置保持一致.

NTU RGB+D包含60个行为类,增加时域扩张残差网络后对每个行为类进行单独的准确度测试,大部分的行为类识别准确率均有提高. 图6和图7是部分常见行为所对应的测试准确度在改进前后的对比,其中包含17个单人行为,3个双人行为.

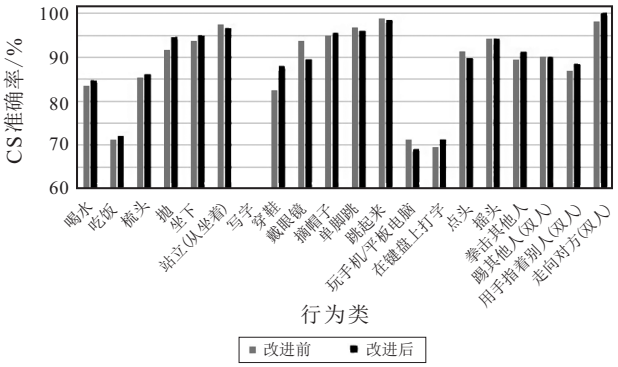

图6 CS准确率改进前后对比

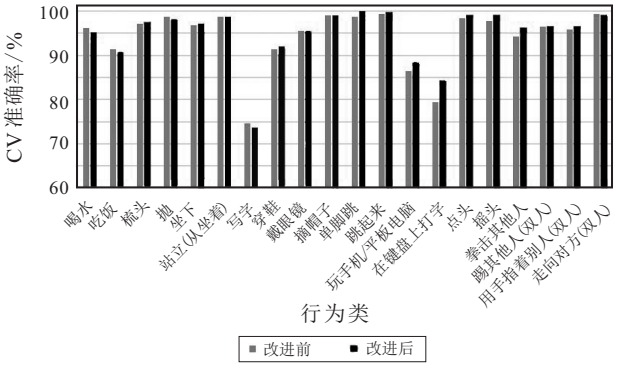

图7 CV准确率改进前后对比

在图6和图7中,改进前是基础主干网络双流自适应图卷积网络[20]行为识别的准确率,改进后是用时域扩张残差网络替换了基础网络中时间卷积网络后的行为识别结果. 可以看出,大部分行为类别的识别准确率都有所提高,这也表明了所构建的时域扩张残差网络的有效性.

在双分支时空行为特征提取中最重要的是确定通道比和帧率比,下面在基于关节点数据的行为识别网络中进行实验. 低帧率分支中的通道数与表1中的相同,高帧率分支为了更好地保证时间保真度没有减少时间帧,即使用原始数据作为输入. 首先将帧率之比设为$a=8$,调整通道数的比值b.

表4为双分支通道数不同比值对应的准确率. 从表4可以看出:当低帧率分支和高帧率分支的通道数之比为8时,CS的准确率最高;当低帧率分支与高帧率分支的通道数之比为4时,CV的准确率最高. 权衡$b=8$与$b=4$的模型计算量,最终将通道比设置为8,测试调整帧率之比的结果如表5所示.

表4 双分支通道数不同比值对应的准确率

通道数比b	CS/%	CV/%
4	88.32	**96.01**
8	**88.91**	95.67
16	87.69	94.98
32	82.97	91.46

表5 双分支帧率不同比值对应的准确率

帧率比a	CS/%	CV/%
4	88.82	95.15
6	**89.02**	**96.29**
8	88.91	95.67
10	83.69	94.98

从表5中可以看出,当双分支帧率比值为6时,CS和CV的行为识别的准确率最高. 在基于骨骼数据的行为识别网络中使用同样的通道数比和帧率比.

为了验证不同改进对最终识别结果的影响,进行消融实验,实验结果如表6所示. 消融实验是以双流自适应图卷积网络(2s-AGCN)[20]作为基础的主干网络进行改进,在该方法中基于两种数据的网络在融合时分类得分直接相加,经实验验证,对不同的网络赋予不同的权重能够使最终特征决策的结果更好,权重调整细节如表3所示.

表6 消融实验

方法	CS/%	CV/%
双流自适应图卷积网络(Baseline)	88.49	95.06
Baseline+加权双流特征决策	88.52	95.13
Baseline+时域扩张残差网络(仅关节点数据)	86.42	93.72
Baseline+时域扩张残差卷网络(仅骨骼数据)	86.13	93.33
Baseline+时域扩张残差网络+加权双流特征决策	**88.71**	**95.62**
Baseline+时域扩张残差网络+双分支结构+加权双流特征决策(ours)	**89.61**	**96.50**

从表6第2行中可以看出,对基础网络仅在特征决策融合时适当调整权重即可使行为识别准确率有一定的提高. 表6第5行中不仅使用了加权双流特征决策,还用时域扩张残差网络替换了基础网络中的时间卷积,使识别准确率有了一定的提升. 从第3、4行中可知,即使将时域扩张残差网络替换了基础网络中的时间卷积,但是,仅使用基于关节点数据的行为

识别网络或仅使用基于骨骼数据的行为识别网络在准确率上均有所下降,这也验证了双流网络的有效性. 第6行中是增加了双分支结构,提高了行为识别结果. 经计算,最终改进的网络参数量与基础网络相比仅增加了5%,说明时域扩张残差网络与双分支结构未使模型复杂度有明显的增加,但是其行为识别准确率得到了提升.

为了表明本文所提出方法的有效性,将实验结果与目前部分先进方法进行对比,见表7. 从表7中可以看出,本文所提出的人体行为识别方法的准确率高于所对比的方法. 在表7的方法中,3scale ResNet152是用多尺度深卷积神经网络进行特征提取的,行为识别准确率较高,但是,这种将关节点数据表示成为图像的方法难以表达不同关节之间的依赖关系. ST-GCN、DPRL+GCNN、AS-GCN这3种方法随着不断地改进,其行为识别准确率有所提高. 2s-AGCN由于将关节点数据与骨骼数据进行互补,使准确率达到了比较优秀的水平. 本文在2s-AGCN的基础上,用时域扩张残差网络代替基础网络中的时间卷积,提高了在时域上的感受野;同时设计了双分支结构,低帧率分支以较少的时间帧数和较多的通道数侧重于提取丰富的空间语义信息,高帧率分支以较多的时间帧数和较少的通道数保证网络轻量级的前提下有效捕捉人体行为的快速变化,提高了行为识别的准确率.

表7　在NTU RGB+D数据集上准确度与先进方法的对比

方法	CS/%	CV/%
3scale ResNet152[8]	85.0	92.3
ST-GCN[14]	81.5	88.3
DPRL+GCNN[18]	83.5	89.8
AS-GCN[19]	86.8	94.2
2s-AGCN[20]	88.5	95.1
ours	**89.6**	**96.5**

4　结　论

将扩张因果卷积与残差连接结构结合构造时域扩张残差网络,能够在未大幅度增加参数量的同时扩大感受野,从而获得在时域上的人体关节信息的长时依赖关系. 同时考虑到在进行分数相加的融合时,由于两个网络所用信息不同而导致性能不同,在进行相加融合时的权重也是不同的,通过进行详细的实验确定了最佳的权重设置结果. 同时对双分支结构中低帧率分支与高帧率分支的帧率之比与通道数之比进行了详细的实验,确定最佳设置比值. 最终使网络训练得到的模型准确率在CS模式上达到89.6%,在CV模式上达到96.5%,高于多数先进行为识别方法,比如ST-GCN[14]、DPRL+GCNN[18]、2s-AGCN[20].

虽然整个网络的准确率较高,但是,训练和测试均是在无任何遮挡的数据集NTU RGB+D上进行的. 在实际应用中,经常会出现人体某些部分被遮挡的情况,目前已有学者对遮挡问题做了一定的研究,但是遮挡状态下行为识别的准确率并不高. 如何提高有遮挡时行为识别的准确率进而提高行为识别网络的鲁棒性将是接下来研究的重点.

参考文献(References)

[1] Ji S W, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[2] Sevillalara L, Liao Y, Guney F, et al. On the integration of optical flow and action recognition[C]. German Conference on Pattern Recognition. Stuttgart: Springer, Cham, 2018: 281-297.

[3] 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.)

[4] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images[C]. CVPR 2011. Colorado Springs, 2011: 1297-1304.

[5] 冉宪宇, 刘凯, 李光, 等. 自适应骨骼中心的人体行为识别算法[J]. 中国图象图形学报, 2018, 23(4): 519-525.
(Ran X Y, Liu K, Li G, et al. Human action recognition algorithm based on adaptive skeleton center[J]. Journal of Image and Graphics, 2018, 23(4): 519-525.)

[6] 刘庭煜, 陆增, 孙毅锋, 等. 基于三维深度卷积神经网络的车间生产行为识别[J]. 计算机集成制造系统, 2020, 26(8): 2143-2156.
(Liu T Y, Lu Z, Sun Y F, et al. Working activity recognition approach based on 3D deep convolutional neural network[J]. Computer Integrated Manufacturing Systems, 2020, 26(8): 2143-2156.)

[7] Shahroudy A, Liu J, Ng T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 1010-1019.

[8] Li B, Dai Y, Cheng X, et al. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN[C]. Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. Hong Kong, 2017: 601-604.

[9] Zhang P F, Lan C L, Xing J L, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]. IEEE International Conference on Computer Vision. Venice, 2017: 2136-2145.

[10] Caetano C, Brémond F, Schwartz W R. Skeleton image representation for 3D action recognition based on tree structure and reference joints[C]. The 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). Rio de Janeiro, 2019: 16-23.

[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]. The 16th IEEE International Conference on Advanced Video and Signal Based Surveillance. Taiwan, 2019: 1-8.

[12] Li Y S, Xia R J, Liu X, et al. Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition[C]. IEEE International Conference on Multimedia and Expo. Shanghai, 2019: 1066-1071.

[13] Yang F, Wu Y, Sakti S, et al. Make skeleton-based action recognition model smaller, faster and better[C]. Proceedings of the ACM Multimedia Asia. Beijing, 2019: 1-6.

[14] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[J/OL]. 2018, arXiv: 1801.07455.

[15] Li C L, Cui Z, Zheng W M, et al. Spatio-temporal graph convolution for skeleton based action recognition[J/OL]. 2018, arXiv: 1802.09834.

[16] 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 (CVPR). Long Beach, 2019: 3590-3598.

[17] Gao X, Hu W, Tang J X, et al. Optimized skeleton-based action recognition via sparsified graph regression[C]. Proceedings of the 27th ACM International Conference on Multimedia. New York, 2019: 601-610.

[18] Tang Y S, Tian Y, Lu J W, et al. Deep progressive reinforcement learning for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, 2018: 5323-5332.

[19] 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 (CVPR). Long Beach, 2019: 3590-3598.

[20] Shi L, Zhang Y F, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 12018-12027.

[21] 练秋生, 富利鹏, 陈书贞, 等. 基于多尺度残差网络的压缩感知重构算法[J]. 自动化学报, 2019, 45(11): 2082-2091.
(Lian Q S, Fu L P, Chen S Z, et al. A compressed sensing algorithm based on multi-scale residual reconstruction network[J]. Acta Automatica Sinica, 2019, 45(11): 2082-2091.)

[22] Fisher Y, Koltun V. Multi-scale context aggregation by dilated convolutions[C]. Proceedings of the IEEE Conference on Learning Representations. San Juan, 2016: 1-15.

[23] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[24] 张小俊, 李辰政, 孙凌宇, 等. 基于改进3D卷积神经网络的行为识别[J]. 计算机集成制造系统, 2019, 25(8): 2000-2006.
(Zhang X J, Li C Z, Sun L Y, et al. Behavior recognition method based on improved 3D convolutional neural network[J]. Computer Integrated Manufacturing Systems, 2019, 25(8): 2000-2006.)

[25] 李倩玉, 蒋建国, 齐美彬. 基于改进深层网络的人脸识别算法[J]. 电子学报, 2017, 45(3): 619-625.
(Li Q Y, Jiang J G, Qi M B. Face recognition algorithm based on improved deep networks[J]. Acta Electronica Sinica, 2017, 45(3): 619-625.)

[26] 宋立飞, 翁理国, 汪凌峰, 等. 多尺度输入3D卷积融合双流模型的行为识别方法[J]. 计算机辅助设计与图形学学报, 2018, 30(11): 2074-2083.
(Song L F, Weng L G, Wang L F, et al. Multi-scale 3D convolution fusion two-stream networks for action recognition[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(11): 2074-2083.)

作者简介

薛盼盼 (1995−), 女, 硕士生, 从事行为识别的研究, E-mail: 15762182757@163.com;

刘云 (1962−), 男, 教授, 博士生导师, 从事计算机视觉、图像处理等研究, E-mail: lyun-1027@163.com;

李辉 (1984−), 男, 副教授, 博士, 从事计算机视觉、行为识别等研究, E-mail: lihui@qust.edu.cn;

陶冶 (1981−), 男, 教授, 博士, 从事机器视觉等研究, E-mail: ye.tao@qust.edu.cn;

田嘉意 (1995−), 女, 硕士生, 从事计算机视觉的研究, E-mail: tianjiayiqust@163.com.

(责任编辑：李君玲)