Journal of Computer Applications
计算机应用,2021,41(7):1915 - 1921
ISSN 1001-9081
CODEN JYIIDU
2021-07-10
http://www.joca.cn

文章编号:1001-9081(2021)07-1915-07
DOI:10.11772/j.issn.1001-9081.2020091515

基于时空注意力图卷积网络模型的人体骨架动作识别算法

李扬志[1],袁家政[2*],刘宏哲[1]
(1.北京市信息服务工程重点实验室(北京联合大学),北京 100101;
2.北京开放大学 科研外事处,北京 100081)
(*通信作者电子邮箱jzyuan@139.com)

摘　要:针对现有的人体骨架动作识别算法不能充分发掘运动的时空特征问题,提出一种基于时空注意力图卷积网络(STA-GCN)模型的人体骨架动作识别算法。该模型包含空间注意力机制和时间注意力机制:空间注意力机制一方面利用光流特征中的瞬时运动信息定位运动显著的空间区域,另一方面在训练过程中引入全局平均池化及辅助分类损失使得该模型可以关注到具有判别力的非运动区域;时间注意力机制则自动地从长时复杂视频中挖掘出具有判别力的时域片段。将这二者融合到统一的图卷积网络(GCN)框架中,实现了端到端的训练。在Kinetics和NTU RGB+D两个公开数据集的对比实验结果表明,基于STA-GCN模型的人体骨架动作识别算法具有很强的鲁棒性与稳定性,与基于时空图卷积网络(ST-GCN)模型的识别算法相比,在Kinetics数据集上的Top-1和Top-5分别提升5.0和4.5个百分点,在NTURGB+D数据集的CS和CV上的Top-1分别提升6.2和6.7个百分点;也优于当前行为识别领域最先进(SOA)方法,如Res-TCN、STA-LSTM和动作-结构图卷积网络(AS-GCN)。结果表示,所提算法可以更好地满足人体行为识别的实际应用需求。

关键词:图卷积网络;人体骨架行为识别;注意力机制;人体关节点;视频行为理解
中图分类号:TP391.4　　**文献标志码**:A

Human skeleton-based action recognition algorithm based on spatiotemporal attention graph convolutional network model

LI Yangzhi[1], YUAN Jiazheng[2*], LIU Hongzhe[1]
(1. *Beijing Key Laboratory of Information Service Engineering* (*Beijing Union University*), *Beijing* 100101, *China*;
2. *Department of Scientific Research and Foreign Affairs*, *Beijing Open University*, *Beijing* 100081, *China*)

Abstract: Aiming at the problem that the existing human skeleton-based action recognition algorithms cannot fully explore the temporal and spatial characteristics of motion, a human skeleton-based action recognition algorithm based on Spatiotemporal Attention Graph Convolutional Network (STA-GCN) model was proposed, which consisted of spatial attention mechanism and temporal attention mechanism. The spatial attention mechanism used the instantaneous motion information of the optical flow features to locate the spatial regions with significant motion on the one hand, and introduced the global average pooling and auxiliary classification loss during the training process to enable the model to focus on the non-motion regions with discriminability ability on the other hand. While the temporal attention mechanism automatically extracted the discriminative time-domain segments from the long-term complex video. Both of spatial and temporal attention mechanisms were integrated into a unified Graph Convolution Network (GCN) framework to enable the end-to-end training. Experimental results on Kinetics and NTU RGB+D datasets show that the proposed algorithm based on STA-GCN has strong robustness and stability, and compared with the benchmark algorithm based on Spatial Temporal Graph Convolutional Network (ST-GCN) model, the Top-1 and Top-5 on Kinetics are improved by 5.0 and 4.5 percentage points, respectively, and the Top-1 on CS and CV of NTU RGB+D dataset are also improved by 6.2 and 6.7 percentage points, respectively; it also outperforms the current State-Of-the-Art (SOA) methods in action recognition, such as Res-TCN (Residue Temporal Convolutional Network), STA-LSTM, and AS-GCN (Actional-Structural Graph Convolutional Network). The results indicate that the proposed algorithm can better meet the practical application requirements of human action recognition.

Key words: Graph Convolutional Network (GCN); human skeleton-based action recognition; attention mechanism; human joint; video behavior understanding

收稿日期:2020-09-29;**修回日期**:2020-12-17;**录用日期**:2020-12-18。
基金项目:国家自然科学资助基金项目(61871028,61871039,61906017,61802019);北京联合大学领军人才项目(BPHR2019AZ01);北京市教委项目(KM202111417001,KM201911417001);北京联合大学研究生科研创新项目(YZ2020K001)。
作者简介:李扬志(1996—),男,湖南邵东人,硕士研究生,主要研究方向:计算机视觉、人工智能;　袁家政(1971—),男,湖南隆回人,教授,博士,主要研究方向:计算机视觉、人工智能;　刘宏哲(1971—),女,河北保定人,教授,博士,主要研究方向:数字图像处理、智能驾驶。

0 引言

行为识别的主要任务是分类识别，对给定的一段动作信息（例如视频、图片、二维骨骼序列、三维骨骼序列），通过特征抽取分类来预测其类别。目前基于视频和RGB（Red，Greeen，Blue）图片的主流方法是双流网络[1]，而基于骨骼数据的主流方法就是图卷积网络（Graph Convolution Network，GCN）。其中卷积神经网络（Convolution Neural Network，CNN）适合提取空间上有相关性的数据，循环神经网络（Recurrent Neural Network，RNN）则适合提取时间上有相关性的数据。现有的基于人工特征或递归神经网络的方法无法充分捕捉骨骼序列复杂的空间结构和长期的时间动态，而这对识别动作非常重要[2]。对基于视频的人体行为识别来说，动作的发生与结束与时间空间紧密相关，并且时间与空间具有潜在关系。然而目前对视频中时空相关数据的提取多依赖于手工设计，耗费了大量人力物力；或者提取的数据不够具有判别性。因此，有效地提取视频数据的时空相关性是解决这些问题的关键。如何探索非线性复杂时空数据，发现其固有的时空模式，准确预测人体行为仍是一个非常有挑战性的课题。

近年来，人体行为识别已经成为了一个活跃的研究领域。一般来说，人体行为的识别可通过多种模式，如外观、深度、光流和人体骨架[3]。其中，动态的人体骨架通常包含有重要信息。动态骨架模态可以表示为一系列人体关节位置的时间序列，从而可通过分析运动模式来对人体行为进行识别。早期Wang等[4]提出利用时间步长的关节坐标形成特征向量来进行骨架动作识别，但没有利用上关节之间的空间关系，所以识别效果有限。Du等[5]提出使用神经网络的方法来学习骨架关节点之间的连接关系。该方法显示了强大的学习能力，并取得了很大的改进，然而大多方法依赖于手工设计的方法或规则，难以泛化。Vemulapalli等[6]提出通过三维骨骼表示来进行人体动作识别，利用人体各部位之间的相对三维几何关系来表示三维人体骨架。近来，基于图的模型[7]因其对图结构数据的有效表示而受到广泛关注。现有的图形模型主要可分为两种架构：图神经网络（Graph Neural Network，GNN）和图卷积网络（Graph Convolutional Network，GCN）。GNN是将图与递归神经网络结合，通过消息传递与节点状态更新的多次迭代，每个节点捕获其相邻节点内的语义关系和结构信息[8]。Qi等[9]提出将GNN应用于图像、视频检测和识别人机交互任务；Li等[10]提出利用GNN对物体之间的依赖关系建模，并预测用于情景识别的一致的结构化输出。GCN则将卷积神经网络扩展到了图模型。Kipf等[11]引入了光谱GCN，用于对图结构数据进行半监督分类；Simonovsky等[12]对空间域内的图信号进行了类似卷积运算，并将图卷积用于点云分类。

基于此，有人提出结合二者优点，利用关节点的自然连接与图模型对图结构物体学习能力的融合，从而更加有效地进行人体行为的检测与识别。Song等[13]提出了时空注意力模型，选择性地关注不同的时空特征，说明时空特征的提取与表示是视频中人体行为识别的关键；Zhang等[14]提出了一种骨架序列的视图自适应模型，能够自动将观测点调整为合适的视点。研究进一步表明，学习区分时空特征是人类动作识别的关键要素。Yan等[15]提出了一种用于动作识别的时空卷积网络（Spatial Temporal Graph Convolutional Network，ST-GCN），每个卷积层用一个图卷积算子构造空间特征，用一个卷积算子建模时空动态。在此基础之上，Li等[16]提出了动作-结构图卷积网络（Actional-Structural Graph Convolutional Network，AS-GCN），通过学习actional-links，扩展structural-links来进行行为识别。此外，Thakkar等[17]提出了一种基于部件的图卷积网络（Part-based Graph Convolutional Network，PB-GCN）来学习部件之间的关系。与ST-GCN和PB-GCN相比，Si等[18]提出利用图神经网络捕获空间结构信息，然后利用长短期记忆（Long Short-Term Memory，LSTM）网络对时间动态进行建模。

然而，动态骨架的建模受到的关注较少，并且缺乏对骨架数据中的空特征的充分发掘，基于此，本文提出了一种时空注意力图卷积网络，目的是发展一种鲁棒的、有效的方法来对动态骨架进行建模，可以自动捕获嵌入在人体骨架关节间的空间和时间动态特征，从而实现人体动作的识别。本文主要工作如下：1）提出一种时空注意力机制来学习人体骨架序列的动态时空相关性，其中空间注意力被用来建模不同节点之间的复杂空间相关性，时间注意力被用来捕捉不同时间之间的动态时间相关性。2）设计了一种新的时空卷积模块，用于骨架数据的时空相关性建模。它由基于图的从骨架网络中获取空间特征的图卷积和描述相邻时间片依赖关系的时间维卷积组成。3）将空间注意力机制和时空注意力机制有机地融合到统一的图卷积中，得到基于时空注意力机制的图卷积网络（Spatiotemporal Attention Graph Convolutional Network，STA-GCN），实现端到端训练，并在骨架动作识别中取得了较好的效果。

1 时空注意力图卷积网络

在人体行为中，人体关节点总是以组为单位进行运动。现有方法已经证明基于骨架的行为识别的有效性[19]，但是现有方法对于时空信息的表示与学习还是依赖手工设计特征或者先验知识，忽略了时间与空间之间的联系，因此本文提出了STA-GCN。

1.1 图卷积网络

骨架是图形的形式，而不是二维或者三维网格，这使得单纯使用卷积网络变得困难。GCN是学习表示图结构数据的一种通用而有效的框架，各种GCN变种已经在许多任务上取得了最先进的成果。

对基于骨架的人体行为识别来说，基于骨架的数据可以从运动捕捉设备或视频中的姿势估计算法中获得。通常数据是一个帧序列，每个帧都会有一组关节坐标。在给定二维或三维坐标形式的人体关节序列的基础上，构造了以关节为图结点、以人体结构和时间的自然连通性为图边的时空图。

图卷积可定义为：

$$f_{\text{out}}(v_{ti}) = \sum_{v_{tj} \in N(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{\text{in}}(v_{tj}) W\left(l_{ti}(v_{tj})\right) \tag{1}$$

其中：f_{in}是节点v_{ti}的特征向量输入；$W(\cdot)$是一个权值函数，并且是根据图标签$l_{ti}: V_t \to \{0, 1, \cdots, K\}$自$K$映射而来，可用于为每个图节点$v_{ti} \in V_t$分配标签，并且可以将节点$v_{ti}$的邻节点集合$N(v_{ti})$划分为固定数量的$K$个子集；$Z_{ti}(v_{tj})$是对特征表示进行规范化的相应子集数量；$f_{\text{out}}\left(v_{ti}\right)$表示图卷积在节点$v_{ti}$处的输出。使用邻接矩阵来代入上式，可得：

$$f_{\text{out}} = \sum_{k=1}^{K} \boldsymbol{\Lambda}_k^{-\frac{1}{2}} \boldsymbol{A}_k \boldsymbol{\Lambda}_k^{-\frac{1}{2}} f_{\text{in}} W_k \tag{2}$$

其中:$\boldsymbol{A}_k$是标签$k \in \{1, 2, \cdots, K\}$在空间构造中的邻接矩阵;$\boldsymbol{\Lambda}_k^{ii} = \sum_j \boldsymbol{A}_k^{ij}$是一个度矩阵。

1.2　时空注意力机制

该模型在图卷积网络中引入了时空注意力机制,可以在长时复杂视频中关注到具判别力的时空区域,同时排除无关区域的干扰。图卷积神经网络的时空注意力机制包含空间注意力机制和时间注意力机制两部分。其中,空间注意力机制一方面利用光流特征中的瞬时运动信息定位视频帧中的运动显著区域,另一方面在训练过程中引入全局平均池化和辅助分类损失使网络关注到具有判别力的非运动区域;而时间注意力机制可以自动地从长时复杂视频中挖掘出最具判别力的视频时域片段,而不需要任何时域标注信息。本文将空间注意力机制和时间注意力机制整合到统一的图卷积神经网络框架中,并实现端到端的训练。

空间注意力网络首先在光流预测数据库上预训练,使得该网络可以关注视频中运动显著的空间区域;然后该网络利用全局平均池化并引入辅助分类损失以增加卷积特征的判别性,以关注到视频中具有判别力的非运动区域;最后,空间注意力网络生成空间注意力热图,该热图突显出空域中运动显著区域以及具有判别力的非运动区域,并指导行为分类网络从感兴趣的空间区域提取有效时空特征,用于行为识别。

按照注意力的可微性,分为硬注意力与软注意力。硬注意力中某个区域要么被关注,要么不关注,这是一个不可微的注意力;软注意力中用0到1的不同分值表示每个区域被关注的程度高低,这是一个可微的注意力。利用空间注意力网络对关键节点进行自适应聚焦,采用软注意力机制自动测量节点的重要性[20]。STA-GCN的中间隐藏层包含丰富的空间结构信息和时间动态信息,有利于关键关节点的选择。

在行为识别的实际应用中,采集的原始视频中总是包含有大量无关或无判别力的视频片段,为排除这些片段的干扰,提出了无监督时间注意力机制。该时间注意力机制不需要任何时域标注信息即可自动地从复杂视频中挖掘出具有判别力的视频时域片段。该机制基于各个视频片段对该视频分类的可信度挖掘视频中具有判别力的时域片段。

2　基于STA-GCN模型的设计实现

基于骨架的数据可以从动作捕捉设备中获取,也可以从视频中利用姿态估计算法获取姿态。通常数据是一系列帧,每一帧都有一组关节坐标。已知人体关节顺序,以人体结构中的关节点为图节点,以时间和关节点的自然连接为边,构造时空图。因此时空注意力图卷积模型的输入是图节点上的联合坐标向量,这可以看作是一种基于图像的CNN,其中输入是由在二维图像网格上的像素位置向量形成的。对输入数据进行多层时空图卷积运算,生成更高层次的特征图,然后通过标准的SoftMax分类器将其分类为相应的动作类别。整个模型采用端到端反向传播的方式进行训练。

STA-GCN模型整体框架如图1所示,在每个时空块中都有时空注意力模块(SAtt、TAtt)和时空卷积模块(GCN、Conv)。为了优化训练效果,在每个组件中采用了残差学习框架[21-22],最后经过行为分类子网络的输出,得到最终预测结果。

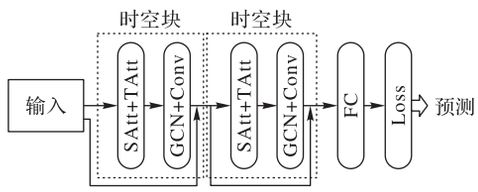


图1　STA-GCN整体框架

Fig. 1　Overall framework for STA-GCN

2.1　时空注意力模块

1)空间注意力。在空间维度上,不同位置的节点状况相互影响,相互影响具有很强的动态性。这里使用注意力机制自适应捕捉空间维度节点之间的动态关联[23]。

$$\boldsymbol{S} = V_s \cdot \sigma((\boldsymbol{x}_h^{(r-1)} w_1) w_2 (w_3 \boldsymbol{x}_h^{(r-1)})^{\mathrm{T}} + b_s) \tag{3}$$

$$S'_{i,j} = \frac{\exp(S_{i,j})}{\sum_{j=1}^{N} \exp(S_{i,j})} \tag{4}$$

其中:$\boldsymbol{x}_h^{(r-1)} = (X_1, X_2, \cdots, X_{T_{r-1}})$表示第$r$个时空块的输入,$C_{r-1}$是第$r$层输入数据的通道数。根据这一层的当前输入动态计算注意力矩阵$\boldsymbol{S}$。T_{r-1}表示时间维度在第r层的长度,V_s, b_s, w_1, w_2, w_3表示学习参数,σ表示激活函数。$\boldsymbol{S}$中一个元素$S_{i,j}$的值语义上表示节点i和节点j之间的关联强度,然后使用Softmax函数保证节点注意力权值的和为1。在执行图卷积时,本文将使用邻接矩阵$\boldsymbol{A}$和空间注意力矩阵$\boldsymbol{S}'$一起来动态调整节点间的影响权值。

2)时间注意力。在时间维度上,不同时间段的人体关节点状况之间存在相关性,且相关性在不同情况下也存在差异。同样地,本文使用注意力机制自适应地对数据给予不同的重视。

$$\boldsymbol{E} = V_e \cdot \sigma(((\boldsymbol{x}_h^{(r-1)})^{\mathrm{T}} u_1) u_2 (u_3 \boldsymbol{x}_h^{(r-1)} + b_e) \tag{5}$$

$$E'_{i,j} = \frac{\exp(E_{i,j})}{\sum_{j=1}^{T_{r-1}} \exp(E_{i,j})} \tag{6}$$

其中:V_e, b_e, u_1, u_2, u_3是学习参数;时间相关矩阵$\boldsymbol{E}$由变化的输入决定。$\boldsymbol{E}$中一个元素$E_{i,j}$的值语义上表示时间i和时间j之间的依赖强度,最后用Softmax函数对$\boldsymbol{E}$进行归一化。直接将归一化时间注意力矩阵应用于输入,使得可以通过$\hat{\boldsymbol{x}}_h^{r-1} = \left(\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_{T_{r-1}}\right) = \left(X_1, X_2, \cdots, X_{T_{r-1}}\right) \boldsymbol{E}'$融合相关信息来动态调整输入。

3)时空注意力融合。空间注意力网络生成空间注意力热图用于指导行为分类网络从感兴趣的空间区域提取有效时空特征;时间注意力机制从原始的复杂视频中自动地挖掘出具有判别力的视频时域片段,并将这些视频片段用于网络训练,而排除其他视频片段对分类器的干扰。

本文提出的时空注意力模型分别在RGB视频帧和光流序列中单独训练两个网络模型,即空间子网络(Spatial Network, SN)和时间子网络(Temporal Network, TN)。然后将两个网络的Softmax预测得分加权融合作为行为分类的依据,如此可以有效提升该网络的分类鲁棒性。需要注意的是,在RGB视频帧数据上训练时,空间注意力网络的参数需要提前在光流预测数据库上预训练,而在光流序列数据上训练时,空间注意力网络与行为分类网络共享权值。

2.2　时空图卷积模块

时空注意力模块使网络自动地对有价值的信息给予相对

较多的注意。将注意力机制调整后的输入输入到时空图卷积模块，其结构如图2所示。本文提出的时空图卷积模块由从邻域中获取空间相关性的空间图卷积和从邻近时间获取时间相关性的时间卷积组成。

谱图理论将网格数据的卷积运算推广到了图形结构数据中。在本研究中，人体骨架网络本质上是一种图结构，每个节点的特征可以看作是图上的信号[24]。因此，为了充分利用人体骨架网络的拓扑特性，本文在每个时间片上采用基于频谱图理论的图卷积直接处理信号，挖掘人体骨架网络在空间维度上的信号相关性。谱方法将图转化为代数形式，分析图的拓扑属性，如图结构中的连通性。

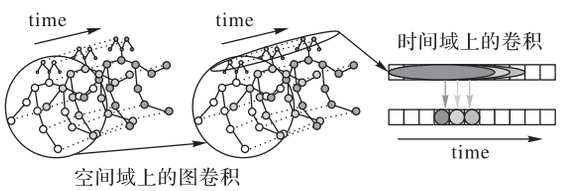


图2 STA-GCN时空图卷积结构

Fig. 2 Spatiotemporal graph convolution structure of STA-GCN

在谱图分析中，图是用对应的拉普拉斯矩阵表示的。通过分析拉普拉斯矩阵及其特征值，可以得到图结构的性质。图的拉普拉斯矩阵定义为：$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$，标准化形式为：$\boldsymbol{L}=\boldsymbol{I}_N-\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\in\mathbf{R}^{N\times N}$。其中：$\boldsymbol{A}$表示邻接矩阵；$\boldsymbol{I}_N$表示单位矩阵；度矩阵$\boldsymbol{D}$是一个对角矩阵，$\boldsymbol{D}_{ii}=\sum_j\boldsymbol{A}_{ij}$。拉普拉斯矩阵的特征值分解为：$\boldsymbol{L}=\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$，其中$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_0,\lambda_1,\cdots,\lambda_{N-1})$为对角矩阵，$\boldsymbol{F}$为傅里叶基。图的卷积是利用在傅里叶域中对角化的线性算子代替经典的卷积算子实现的卷积运算[25]。基于此，图上的信号x可通过核g_θ变换进行滤波：

$$g_\theta *_{\mathrm{G}} x=g_\theta(\boldsymbol{L})x=g_\theta(\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^{\mathrm{T}})x=\boldsymbol{F}g_\theta(\boldsymbol{\Lambda})\boldsymbol{F}^{\mathrm{T}}x \tag{7}$$

其中$*_{\mathrm{G}}$表示图卷积操作。由于图形信号的卷积运算等于通过图形傅里叶变换将这些信号变换到谱域的乘积，上式可以理解为将g_θ的极小部分和x分别在谱域进行傅里叶变换，然后将变换后的结果相乘，再进行傅里叶反变换，得到卷积运算的最终结果。但是，当图的尺度较大时，直接对拉普拉斯矩阵进行特征值分解的代价较大，因此，本文采用切比雪夫多项式近似而有效地解决该问题：

$$g_\theta *_{\mathrm{G}} x=g_\theta(\boldsymbol{L})x=\sum_{k=0}^{K-1}\boldsymbol{\gamma}_k T_k(\tilde{\boldsymbol{L}})x \tag{8}$$

其中：$\boldsymbol{\gamma}$表示多项式系数的一个向量；$\tilde{\boldsymbol{L}}=\frac{2}{\lambda_{\max}}\boldsymbol{L}-\boldsymbol{I}_N$，$\lambda_{\max}$是拉普拉斯矩阵的最大特征值。切比雪夫多项式的递归定义为$T_k(x)=2xT_{k-1}(x)-T_{k-2}(x)$，其中：$T_0(x)=1$，$T_1(x)=x$。利用切比雪夫多项式的近似展开求解，对应于通过卷积核g_θ来提取一图中以每个节点为中心到第$(K-1)$阶邻域信息。图的卷积模块使用经过校正的线性单元作为最终的激活函数。

在对图进行卷积运算后，获取图上每个节点在空间维度上的邻近信息，在时间维度上进一步堆叠标准卷积层，通过合并相邻时间片上的信息来更新节点的信号。以最邻矩阵在第r层的操作为例：

$$\boldsymbol{x}_h^{(r)}=\mathrm{ReLU}(Q*(\mathrm{ReLU}(g_\theta *_{\mathrm{G}}\hat{\boldsymbol{x}}_h^{(r-1)})))\in\mathbf{R}^{C_r\times N\times T_r} \tag{9}$$

其中：$*$表示标准卷积操作，Q表示时间维卷积核的参数，ReLU()表示激活函数。

综上所述，时空卷积模块能够很好地捕捉人体行为数据的时空特征。一个时空注意力模块和一个时空卷积模块构成一个时空块，将多个时空块叠加，可进一步提取更大范围的动态时空相关性。最后，附加一个全连接层，保证各分量的输出与预测目标具有相同的维数和形状。最后的全连接层使用ReLU作为激活函数。

3 实验与分析

为测试时空注意力图卷积网络在基于骨架的动作识别数据集上的表现，选用以下两个数据集：1）Kinetics human action dataset[26]，它是目前为止最大的无约束动作识别数据集；2）NTU RGB+D[27]，它是目前最大的室内动作识别数据集。为了检测本文模型对识别性能的贡献，首先在动作识别数据集上进行了详细的消融实验。为了验证时空注意力图卷积网络是否有效，本文将识别结果与其他最先进的方法进行了比较。所有实验都在Pytorch深度学习框架上进行，并使用了8个TITANX GPUs。

3.1 数据集及评价指标

Kinetics数据集包含了从YouTube上检索到的大约300 000个视频剪辑。这些视频覆盖了多达400类人体行为，主要有日常活动、体育、复杂交互动作等；每个视频片段持续10 s左右。

该数据集只提供了原始视频剪辑，没有骨架数据。由于本文主要是针对基于骨架的动作识别，因此需要将原始帧转换成关节位置序列，主要借用了OpenPose[28]工具箱来进行关节位置的捕获。为了获取关节位置，统一将视频分辨率更改为340×256，并将帧率转换为30帧每秒（Frames Per Second，FPS）。工具箱给出了像素坐标系中的二维坐标(X,Y)和18个人体关节的置信度得分D，故可用(X,Y,D)的元组表示每个关节，一个人体骨架被记录为18个元组的数组。对于多人情况，选取视频剪辑中平均置信度最高的2人。通过这种方式，一个带有T帧的剪辑被转成这些元组的骨架序列。用$(3,T,18,2)$维张量表示剪辑，为简化处理，设置$T=300$。

本文根据数据集作者推荐的Top-1和Top-5分类精度来评估识别性能，将数据集划分为240 000个剪辑的训练集和20 000个剪辑的验证集。

NTU RGB+D数据集是目前最大的人体行为识别3D关节注释数据集，共包含60个动作类的56 000个动作剪辑。这些短片由40位志愿者在一个实验室由3个摄像机拍摄。所提供的注释给出了由深度传感器Kinect检测到的关节点3D坐标位置(X,Y,Z)，其中每个受试者的骨架序列中含有25个关节，每个短片中最多有两名受试者。

根据作者推荐，该数据集可划分为两类：1）CS（Cross-Subject）：包含40 320个剪辑的训练集与16 560个剪辑的测试集。在这种划分下，一部分志愿者只出现在训练集，一部分只出现在测试集。2）CV（Cross-View）：包含37 920个剪辑的验证集与18 960个剪辑的测试集。这种划分下，用于训练的剪辑来自摄像头2和3，测试集的剪辑来源于摄像头1。本文将遵循这个惯例，并在两个划分上以Top-1分类精度作为识别性能评估指标。

3.2 消融实验

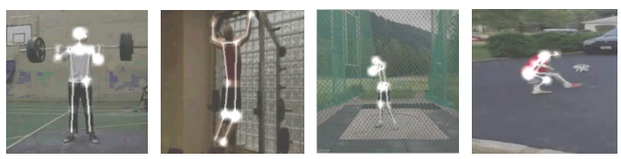
本节共设计了四组实验，详细验证本文提出的基于时空

注意力机制的图卷积神经网络在骨架动作识别任务中的有效性。第一组实验展示了注意力模型网络学习的对应帧的注意力热图的可视化结果;第二、三组实验为切片实验,用于单独验证空间注意力机制和时间注意力机制的有效性;第四组实验是将空间注意力机制和时间注意力机制融合到一个图卷积神经网络中,实现端到端的训练,并将其应用到动作识别任务中。该实验是为了验证在图卷积神经网络中同时引入时空注意力机制对动作识别任务的提升效果。

在第一组实验中,在视频序列的每个时刻根据当前时刻的输入特征和记忆的历史信息分别生成显著性热力图。

如图3所示,该热力图显示了不同关节点位置的重要程度。根据热力图可以发现,不同动作对关节的关注程度也有所区别,对于示例(a)来说,其运动显著区域是手及杠铃周围,因此对这些部位的关注程度要高。

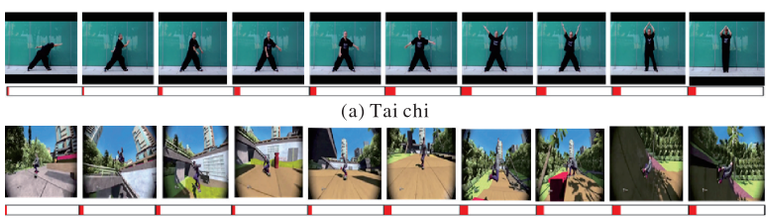
图4展示了示例视频各时域片段及其预测可信度。其中每一个视频示例上方的一行展示的是视频各个时域片段的关键帧,下方的红色长条表示不同视频时域片段对应的置信度,该置信度由本文提出的时间注意力机制学习得到。



(a) Clean and jerk　(b) Pull up　(c) Hammer throw　(d) Juggling ball

图3　空间注意力热力图

Fig. 3　Heat maps of spatial attention



(a) Tai chi

(b) Skateboarding

图4　视频时域片段及其预测置信度

Fig. 4　Video time-domain segments and their prediction confidences

图5展示了利用时空注意力模型得到的视频帧注意力热力图及置信度。其中,视频帧中人体关节位置的颜色深度代表不同的重要程度,颜色越深代表该关节受重视的权重越高。而视频帧下方的红条代表该帧在视频片段的置信度,红色越长,代表该片段对于识别的判别性越好。

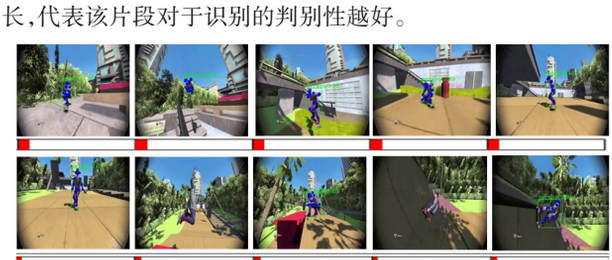


图5　视频帧时空注意热力图及置信度

Fig. 5.　Spatiotemporal attention heat maps and confidences of video frames

第二组实验主要是在Kinetics和NTU RGB+D动作识别数据集上验证空间注意力机制的有效性。在该组实验中,本文比较了基于空间注意力机制的图卷积模型与对应的图卷积网络基准模型的识别精度。ST-GCN模型在骨架动作识别任务中取得了较好结果,本实验以ST-GCN模型为基础,在该网络模型中引入空间注意力机制(记作SA),以验证空间注意力机制在动作识别任务中的有效性。

如表1所示,基于空间注意力机制的行为识别模型的识别精度明显优于ST-GCN基准模型。

第三组实验主要是在动作识别数据集上验证时间注意力机制的有效性。本组实验共设计了两种个不同的基于时间注意力机制的卷积模型,均是在基准模型ST-GCN基础之上加入时间注意力机制(记作TA)。时间注意力机制卷积模型采用不同的视频时域分割及选择方式,以验证不同时域分割方式及不同分割参数对行为识别精度的影响。在该组实验中基准模型的训练过程与上一组实验相同。实验结果如表2所示,其中,平均融合方式(Average Fusion Mode, AFM)是将各个时域片段的预测结果直接平均后作为整个视频的预测结果;区分信任加权(Discriminative Confidence Weighting, DCW)表示基于预测可信度加权的时域融合方式,即根据各个输入片段的预测可信度选择其中最可靠视频片段,并将这些片段的预测结果进行加权作为整个视频的预测结果。

表1　验证空间注意力机制在动作识别的有效性

Tab. 1　Verification of effectiveness of spatial attention mechanism in action recognition

模型	Top-1/%			Top-5/%
	Kinetics	NTU RGB+D		Kinetics
		CS	CV	
ST-GCN[12]	30.7	81.5	88.3	52.8
SA+ST-GCN	**33.1**	**86.3**	**93.7**	**54.9**

表2　验证时间注意力机制在动作识别中的有效性

Tab. 2　Verification of effectiveness of temporal attention mechanism in action recognition

模型	Top-1/%			Top-5/%
	Kinetics	NTU RGB+D		Kinetics
		CS	CV	
ST-GCN[12]	30.7	52.8	81.5	88.3
TA+ST-GCN (AFM)	**33.5**	**54.7**	**85.3**	**91.4**
TA+ST-GCN (DCW)	**34.3**	**55.6**	**85.1**	**92.6**

从表2的实验结果分析可得出如下两个结论:1)基于时间注意力机制的行为识别模型的分类精度均优于基准模型ST-GCN。在测试中使用DCW时域融合方式,在Kinetics数据集上测试时,基于时间注意力机制的模型比基准模型在Top-1和Top-5上分别提高了3.6和2.8个百分点。这些提升说明了本文提出的时间注意力机制在行为识别任务中的有效性。2)对基于时间注意力机制的行为识别模型而言,在测试中使

用DCW时域融合方法的分类精度明显高于使用AFM时域融合方法得到的分类精度。该实验表明,在训练中使用时间注意力机制,测试中DCW的时域融合方法比简单平均融合方法更有效。

第四组实验用于验证时空注意力机制在动作识别任务中的效果。本组实验将空间注意力机制与时间注意力机制融合到统一的图卷积神经网络框架中,得到基于时空注意力机制的图卷积神经网络,并实现端到端训练。其中ST-GCN是基准模型,该模型的训练方法与前两个实验相同,但是在该组实验中基准模型融合SN和TN两个网络模型的预测分布以提升单个模型的分类精度。表3列举了基于时空注意力机制的行为识别模型和基准模型的分类结果。

从表3数据可以看出,在图卷积网络中同时引入时时间注意力机制与空间注意力机制对提升动作识别精度非常有效,且比单独引入空间注意力机制或者时间注意力机制时识别精度提升得更加明显。

表3 验证时空注意力机制在动作识别中的有效性

Tab. 3 Verification of effectiveness of spatiotemporal attention mechanism in action recognition

模型	Top-1/%			Top-5/%
	Kinetics	NTU RGB+D		Kinetics
		CS	CV	
ST-GCN[12]	30.7	52.8	81.5	88.3
STA-GCN(本文模型)	**35.7**	**57.3**	**87.7**	**95.0**

3.3 先进方法实验比较

为了验证STA-GCN在人体骨架动作识别任务中的可竞争性性,将基于时空注意力机制的图卷积网络训练的模型与当前动作识别领域最先进(State-Of-the-Art, SOA)方法进行对比,包括:1) Res-TCN (Residue Temporal Convolutional Network)[29],该方法通过重新构造具有剩余连接的时域卷积(Temporal Convolutional Network, TCN)来提高模型的可解释性。2)STA-LSTM[10],该方法在具有长短时记忆的递归神经网络的基础上建立时空注意力LSTM网络模型,可以选择性关注输入帧的关节差异,并对不同帧的输出给予不同程度的关注,因此可以提取具有区分性的时空特征帮助动作识别。3)ST-GCN[12],它突破了以往骨骼建模方法的局限性,将图卷积应用于人体骨架动作识别,并且提出的模型具有较强的泛化能力。4)AS-GCN[13],该方法通过将A-links及S-links结合成一个广义骨架图,进一步建立行为结构图卷积网络模型来学习空间和时序特征,能够更准确详细地捕捉不同动作模式。实验结果如表4所示。

表4 STA-GCN与当前骨架动作识别领域最先进方法的比较

Tab. 4 Comparison between STA-GCN with current state-of-the-art methods in field of skeleton-based action recognition

模型	Top-1/%			Top-5/%
	Kinetics	NTU RGB+D		Kinetics
		CS	CV	
Res-TCN[29]	20.3	40.0	—	—
STA-LSTM[10]	—	—	73.4	81.2
ST-GCN[12]	30.7	52.8	81.5	88.3
AS-GCN[13]	34.8	56.5	86.8	94.2
STA-GCN	**35.7**	**57.3**	**87.7**	**95.0**

根据表4的实验结果分析,可以得出以下结论:1)由于引入了有效的时空注意模型与训练策略,能够提取具有判别力的时空特征,本文提出的STA-GCN在这两个数据集上均获得了当前同类方法中最好的分类精度;2)通过与Res-TCN[29]模型比较,图卷积相比传统卷积网络更适合基于骨架的动作识别;3)与其他基于LSTM(STA-LSTM[10]和AS-GCN[13])的网络模型相比,本文提出的时空注意力机制模型不仅能够有效捕获骨架数据的时间特征,并且识别性能更加优越。

4 结语

本文提出了基于时空注意力图卷积网络(STA-GCN)模型的人体骨架动作识别算法,STA-GCN模型在骨架序列上构造了一组时空图卷积,提出的时空注意力机制可以同时捕捉空间构造和时间动态的判别特征,而且可以探索时空域之间的关系。在两个具有挑战性的大型数据集Kinetics和NTURGB+D上与目前具有代表性的SOA方法进行了对比,结果显示该模型能获得最优的结果。STA-GCN的灵活性也为未来的工作开辟了许多可能方向,例如,如何将场景、物体和交互等上下文结合,以提升识别性能。

参考文献 (References)

[1] PENG Y, ZHAO Y, ZHANG J. Two-stream collaborative learning with spatial-temporal attention for video classification [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(3):773-786.

[2] KE Q, BENNAMOUN M, AN S, et al. Learning clip representations for skeleton-based 3D action recognition [J]. IEEE Transactions on Image Processing, 2018, 27(6): 2842-2855.

[3] CHEN C, LIU K, KEHTARNAVAZ N. Real-time human action recognition based on depth motion maps [J]. Journal of Real-Time Image Processing, 2016, 12(1):155-163.

[4] WANG J, LIU Z, WU Y, et al. Mining actionlet ensemble for action recognition with depth cameras [C]// Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 1290-1297.

[5] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1110-1118.

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

[7] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks? [EB/OL]. [2020-01-07]. https://arxiv. org/pdf/1810. 00826. pdf.

[8] ZHANG M, CHEN Y. Link prediction based on graph neural networks [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018: 5171-5181.

[9] QI S, WANG W, JIA B, et al. Learning human-object interactions by graph parsing neural networks [C]// Proceedings of the 2018 European Conference on Computer Vision, LNCS 11213. Cham: Springer, 2018: 407-423.

[10] LI R, TAPASWI M, LIAO R, et al. Situation recognition with graph neural networks [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 4183-4192.

[11] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks [EB/OL]. [2020-01-17]. https://arxiv. org/pdf/1609. 02907. pdf.

[12] SIMONOVSKY M, KOMODAKIS N. Dynamic edge-conditioned filters in convolutional neural networks on graphs[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern recognition. Piscataway: IEEE, 2017: 29-38.

[13] 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 31st AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 4263-4270.

[14] ZHANG P, LAN C, XING J, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2136-2145.

[15] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition [EB/OL]. [2020-02-13]. https://arxiv. org/pdf/1801. 07455. pdf.

[16] LI M, CHEN S, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3590-3698.

[17] THAKKAR K, NARAYANAN P J. Part-based graph convolutional network for action recognition [C]// Proceedings of the 2018 British Machine Vision Conference. Durham: BMVA Press, 2018: No. 1003.

[18] SI C, JING Y, WANG W, et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning[C]// Proceedings of the 2018 European Conference on Computer Vision, LNCS 11205. Cham: Springer, 2018: 106-121.

[19] ZHANG S, LIU X, XIAO J. On geometric features for skeleton-based action recognition using multilayer LSTM networks [C]// Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2017: 148-157.

[20] ZHANG P, LAN C, XING J, et al. View adaptive neural networks for high performance skeleton-based human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1963-1978.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[22] CAO C, LAN C, ZHANG Y, et al. Skeleton-based action recognition with gated convolutional neural networks [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(11): 3247-3257.

[23] FENG X, GUO J, QIN B, et al. Effective deep memory networks for distant supervised relation extraction [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 4002-4008.

[24] SHUMAN D I, NARANG S K, FROSSARD P, et al. The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains [J]. IEEE Signal Processing Magazine, 2013, 30(3):83-98.

[25] HENAFF M, BRUNA J. LeCUN Y. Deep convolutional networks on graph-structured data [EB/OL]. [2020-03-10]. https://arxiv. org/pdf/1506. 05163. pdf.

[26] KAY W, CARREIRA J, SIMONYAN K, et al. The Kinetics human action video dataset[EB/OL]. [2020-03-10]. https://arxiv. org/pdf/1705. 06950. pdf.

[27] 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 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1010-1019.

[28] 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.

[29] KIM T S, REITER A. Interpretable 3D human action analysis with temporal convolutional networks[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2017:1623-1631.

[30] 朱大勇,郭星,吴建国. 基于Kinect三维骨骼节点的动作识别方法[J]. 计算机工程与应用, 2018, 54(20):152-158. (ZHU D Y, GUO X, WU J G. Action recognition method using Kinect 3D skeleton data[J]. Computer Engineering and Applications, 2018, 54(20):152-158.)

[31] 喻露,胡剑锋,姚磊岳. 基于人体骨架的非标准深蹲姿势检测方法[J]. 计算机应用, 2019, 39(5):1448-1452. (YU L, HU J F, YAO L Y. Detection method of non-standard deep squat posture based on human skeleton [J]. Journal of Computer Applications, 2019, 39(5):1448-1452.)

This work is partially supported by the National Natural Science Foundation of China (61871028, 61871039, 61906017, 61802019), the Leading Talent Program of Beijing Union University of China (BPHR2019AZ01), the Beijing Municipal Education Commission Project of China (KM202111417001, KM201911417001), the Graduate Research and Innovation Funding Project of Beijing Union University (YZ2020K001)。

LI Yangzhi, born in 1996, M. S. candidate. His research interests include computer vision, artificial intelligence.

YUAN Jiazheng, born in 1971, Ph. D., professor. His research interests include computer vision, artificial intelligence.

LIU Hongzhe, born in 1971, Ph. D., professor. Her research interests include digital image processing, intelligent driving.