DOI:10.19651/j.cnki.emt.2108617

基于残差融合双流图卷积网络的手势识别方法*

程换新[1] 成 凯[1] 程 力[2] 蒋泽芹[1]
(1.青岛科技大学自动化与电子工程学院 青岛 266061;2.中国科学院新疆理化技术研究所 乌鲁木齐 830011)

摘 要: 针对传统图卷积网络易忽略空间特征与时间特征之间关联的问题,设计了一种基于残差结构和图卷积网络相融合的双流网络模型。首先网络包括空间流和时间流两个通道,将手势骨骼数据构建成空间图和时序图作为两通道的输入,通过分离时间维度和空间维度极大地提高了训练速度。然后为了增加网络深度,避免梯度消失等问题,嵌入残差结构并对其进行改进,更加有效利用时间特征,保证了特征的多样性。最后将两通道输出的空间点集序列和时间边集序列串联转化,输入 Softmax 分类器进行分类,得到识别结果。将新提出的方法在 CSL 和 DEVISIGN-L 手势数据集上进行实验,结果表明在两个数据集上识别精度分别达到了 96.2%和 69.3%,证明该方法具有一定的先进性。

关键词: 手势识别;残差结构;双流图卷积网络
中图分类号: TP391.9 **文献标识码:** A **国家标准学科分类代码:** 520.60

Gesture recognition method based on residual fusion dual-stream graph convolutional network

Cheng Huanxin[1] Cheng Kai[1] Cheng Li[2] Jiang Zeqin[1]
(1. College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China;
2. Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China)

Abstract: Aiming at the problem that the traditional graph convolutional network ignores the relationship between spatial and temporal features, a dual-stream network model based on the combination of residual structure and graph convolutional network is proposed. First of all, the network includes two channels of space flow and time flow. The gesture skeleton information is constructed into a space diagram and a time sequence diagram as the input of the two channels. The training speed is greatly improved by separating the time dimension and the space dimension. Then, in order to increase the depth of the network and avoid problems such as the disappearance of gradients, the residual structure is embedded and improved to make more effective use of time features and ensure the diversity of features. Finally, the spatial point set sequence and the time edge set sequence output by the two channels are converted in series and input into the Softmax classifier for classification, and the recognition result is obtained. The newly proposed method is tested on the CSL and DEVISIGN-L gesture datasets, and the results show that the recognition accuracy on the two datasets reaches 96.2% and 69.3%, which proves that the method has a certain degree of advancement.

Keywords: gesture recognition; residual structure; two-stream graph convolutional network

0 引 言

随着科学技术的不断发展,人与计算机之间的相互联系持续增强,人机交互已经从原始计算机作为主体发展到以人为主体,交互方法也逐渐丰富。手势作为一种直观生动的行为方式,包含各种丰富而重要的信息。手势通常由一个或多个手部动作相互转换组成,其中一个动作的变化可能会导致另外一种完全不同的含义。手势识别的应用领域广泛,主要包括手语识别[1],人类手部动作识别[2],姿势检测,体育运动检测,医疗辅助等方面,具有十分重要的研究价值。

根据数据模式的不同,手势识别可以分为基于传感器的识别方法和基于计算机视觉的识别方法。在手势识别的早期阶段,主要依赖于带有传感器的数据手套。文献[3]设计了一种可穿戴传感器作为输入设备的手势识别系统,能够识别简单手势。文献[4]提出了一种基于三轴加速度计

收稿日期:2021-12-17
* 基金项目:国家海洋局重大专项(国海科字[2016]194 号 No.30)资助

多通道肌电图传感器的框架用于手势识别,达到了较高的识别精度。基于传感器的方法具有识别速度快和精度高的特点,但电子设备的穿戴通常会带来很多不便,具有一定的局限性。

随着深度学习技术的迅猛发展,在计算机视觉领域中的作用越来越大,为手势识别提供了一种新的识别方向。卷积神经网络基于其强大的特征提取能力,在手势识别领域占据了重要地位。Cui 等[5]开发了一种基于计算机视觉的连续手势识别框架,采用具有堆叠时间融合层的深度卷积神经网络作为特征提取模块,双向循环神经网络作为序列学习模块,取得了不错的识别效果。Pigou 等[6]提出了一种新的端到端可训练神经网络架构,结合了时间卷积和双向递归,探索了视频中手势识别的深层架构,取得先进的识别效果。使用 CNN 网络进行特征提取可以提高网络训练效率,为了不遗漏视频的时空信息,文献[7]提出了一种 3D 卷积神经网络进行动态手语识别。该网络使用高采样频率分支关注图像中的运动信息,使用低采样频率分支关注语义信息,然后融合两个分支特征完成手语识别。与 CNN 相比,RNN 更加适合处理序列数据,捕捉长期的上下文语义信息。Huang 等[8]提出了一种基于以关键帧为中心的新型序列到序列的网络结构,将关键帧算法嵌入到 RNN 网络中,保证手势中关键信息的完整性,从而取得了良好的识别效果。Masood 等[9]提出了一种使用 CNN 和 RNN 从视频序列中识别手势的方法,使用梯度下降反向传播的方式训练神经网络,首先利用卷积神经网络提取图像特征,然后生成特征向量并输入 RNN 进行计算。这些方法虽然取得了一定的识别效果,但 RNN 和 CNN 都不能完全代表骨骼数据的结构,无法准确的表示出骨骼关节坐标与手势之间的关系,泛化能力较差。

基于骨骼数据的手势识别方法在动作表示有效性强,骨骼数据更加稀疏和稳定,并且与动作的基本结构相一致。在手势识别过程中,手势通过手的骨骼数据表示,然而 CNN 和 RNN 无法有效的建模骨骼数据。随着图卷积神经网络(GCN)[10]高速发展,在动作识别领域得到了广泛的应用。GCN 是一种通用的深度学习框架,可以直接应用于结构化数据,同时骨骼数据是自然的图形数据,非常适合使用 GCN 算法进行建模。Meng 等[11]提出了一种基于图卷积网络的多尺度手语识别网络(SLR-Net),首先从视频中提取骨骼数据,然后将骨骼数据用于手势识别。该网络克服了动作模糊、手势风格多样的难点,取得了良好的识别效果。Yan 等[12]提出了一种新的动态骨骼模型,称为时空图卷积网络(ST-GCN),通过从数据中自动学习空间和时间模式,超越了以前方法的局限性,提高了表达能力和泛化能力。Xiong 等[13]提出了一种基于手势骨骼提取器和多通道扩张图卷积网络的交警手势识别网络,通过提取具有判别性和可解释性的手势骨骼坐标信息,解决了复杂背景和噪声而导致的识别不准确问题。

残差网络结构在图像分类、图像修复等视觉任务中有着显著效果,为了增加网络模型的深度并避免梯度消失等问题,通常会使用残差网络结构对其进行改进。文献[14]提出了一种基于注意力机制的残差 3D 卷积网络模型用于人体动作识别,利用残差网络学习视频序列中的时间相关性,从而避免梯度消失问题,取得了优秀的识别效果。Liao 等[15]提出了一种基于深度三维残差和双向 LSTM 网络的多模态动态识别网络,网络从视频序列中提取时空特征,并在特征分析后建立与视频序列中每个动作对应的中间分数,然后加入残差结构减少了网络计算时间和空间复杂度,提高识别精度。

本文提出了一种基于残差结构和图卷积网络相结合的双流网络模型。该网络将手势骨骼信息构建成空间图和时序图作为空间流和时间流通道的输入,通过图卷积提取手部姿势特征,利用卷积运算对这些特征进行下采样,保持骨骼关节坐标信息和手势之间的联系。加入改进残差结构增加网络深度,加快网络收敛速度,同时避免梯度消失等问题。利用残差结构将提取的高维特征分成几个具有相同维度的浅层特征,保证了特征的多样性。最后将两通道输出的时间边集序列和空间点集序列串联转化,输入 Softmax 分类器进行分类,得到识别结果。在 CSL 动态手势数据集上实验结果表明,该方法具有一定的先进性

1 方法设计

1.1 图卷积运算

图卷积网络常用于处理基于图结构数据的工作中,非常适合表示骨骼特征。由于关节具有自然的结构和连接,因此骨骼数据可以用一个图表示,其中每个顶点表示一个关节。

本文使用图卷积神经网络来挖掘手部关节之间的空间关系,建模骨骼数据。把骨骼的关节点表示为三维矢量,分别对应 x、y、z 坐标。在给定的一定时间内两个相邻连接点为 $v_1=(x_1,y_1,z_1)$ 和 $v_2=(x_2,y_2,z_2)$。定义骨骼数据对应的关节点集合为 $\boldsymbol{V}=\{v_i\}_{i=1}^{n}$,骨骼数据集合为 $\boldsymbol{W}=\{w_i\}_{i=1}^{n}$,其中 $w=v_1-v_2$。此时,骨骼数据图可以表示为 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{W})$。其中 $\boldsymbol{V}$ 作为一个矩阵可表示为 $\boldsymbol{X}=(v_1,v_2,\cdots,v_n)$,$v_i\in\mathbb{R}^c$。$\boldsymbol{A}\in\mathbb{R}^{n\times n}$ 表示 $\boldsymbol{X}$ 的邻接矩阵,$\boldsymbol{D}\in\mathbb{R}^{n\times n}$ 表示次数矩阵,符合条件 $D_{ij}=\sum_i A_{ij}$。g 表示过滤器,是 $\boldsymbol{B}\in\mathbb{R}^{c\times r}$ 的关键参数。因此图的卷积运算可以定义为:

$$g_\theta\otimes\boldsymbol{X}=(\boldsymbol{I}+\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B} \tag{1}$$

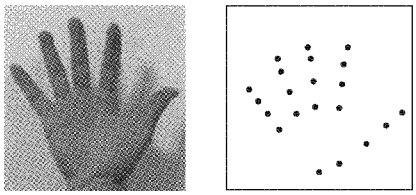
式中:$\otimes$是图卷积操作,通过该操作可避免梯度爆炸等问题。使用 $\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$ 和 $\hat{\boldsymbol{D}}=\sum_i\hat{A}_{ij}$ 代入式(1)中,图卷积运算可以仅一步表示为:

$$g_\theta\otimes\boldsymbol{X}=\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}-\frac{1}{2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B} \tag{2}$$

1.2　构建空间图和时序图

原始骨骼数据通常是从一系列视频帧中提取出来的，或由特定采集设备采集获取。这种骨骼数据包含一系列节点信息，需要进一步将其表示为一个图数据结构，以便定义卷积运算并构造卷积块。

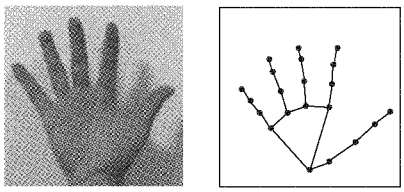
手势的空间图可以表示为 $\partial=(\boldsymbol{V},\boldsymbol{X},\boldsymbol{A})$，如图 1 所示，其中 $\boldsymbol{V}=(v_1,v_2,\cdots,v_N)$ 表示节点集，$\boldsymbol{A}\in\{0,1\}^{N\times N}$ 表示邻接矩阵。如果有从 v_i 到 v_j 的边，则 $A_{ij}=1$，否则 $A_{ij}=0$。$\boldsymbol{X}\in\mathbb{R}^{N\times V\times M}$ 可以表示每一个节点的特征，d 表示表示特征通道的数量，N 表示节点的数量。其中 N 代表帧数，V 代表每一帧的关节数，M 代表通道数。



(a) 手势图像　(b) 骨骼数据空间图

图 1　构建空间图

由于人的手势是有序的，需要构建另一个时序图来捕获时间特征。当用 $\boldsymbol{X}\in\mathbb{R}^{N\times V\times M}$ 表示骨骼数据的节点特征时，第 i 帧骨骼数据是 $X_i\in\mathbb{R}^{V\times M}$，如图 2 所示。由此可以从 X_i 构建骨骼数据图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{W})$，其中，$\boldsymbol{V}$ 作为定点的集合矩阵可表示为 $\boldsymbol{X}=(v_1,v_2,\cdots,v_n)$，$\boldsymbol{W}$ 是连接骨骼数据图中任意两个顶点的边的集合。本文将手势的关节作为顶点，将骨骼设置为边，从而将骨骼数据构建为时序图形。



(a) 第 i 帧手势图像　(b) 骨骼数据时序图

图 2　构建时序图

1.3　改进残差结构

本文使用残差结构(ResNet)加快网络的收敛速度，提高识别精度，避免梯度消失问题。残差结构如图 3 所示，F 代表残差单元，$F(x)+x$ 代表期望输出，2 个通道内核大小为 3×3。对于 l 层中的每个单元可进行如下表示：

$$x_{l+1}=x_l+F(x_l+W_l) \tag{3}$$

式中：x_l 代表输入，W_l 代表学习参数，激活函数采用 ReLU 函数。网络模型的卷积层对骨骼数据图进行输入操作，将生成的激活映射函数传递给后续层。

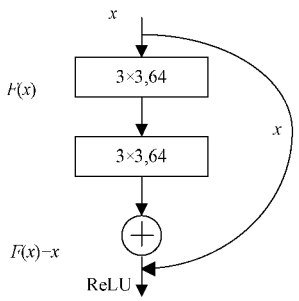


图 3　残差结构

为了更加有效利用提取的时间特征，需要对残差结构进行改进，称为 2D-ResNet 结构，如图 4 所示。该结构利用二维卷积提取时间特征，然后扩张并重构特征，使用 6×1 大小的滤波器提取序列时间信息。

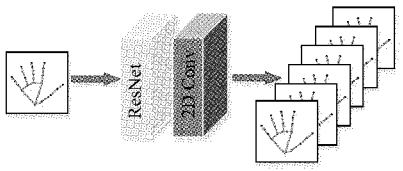


图 4　2D-ResNet 结构

1.4　网络模型

本文将残差结构嵌入到图卷积网络中设计了一种空间流和时间流的新型双流网络，称为 SCR-GAN，该网络通过分离时间维度和空间维度极大地提高了训练速度，如图 5 所示。

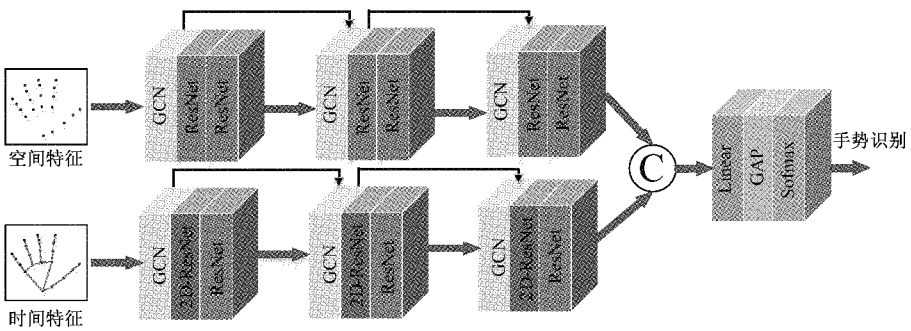


图 5　网络模型

网络的两个通道都包括 3 个编码结构。空间流通道的输入为骨骼数据空间图，每个编码结构包括 1 个图卷积编码块和 2 个 ResNet 结构。该设计可以将输入骨骼数据转化为空间更小、通道更多的特征映射，从而有效地维持网络的接受域。时间流通道的输入为骨骼数据时序图，每个编码结构包括 1 个图卷积编码块、1 个 2D-ResNet 结构

和 1 个 ResNet 结构,其中 2D-ResNet 结构能更加有效利用提取的时间特征。3 个编码结构的输出通道数分别为 64、128、256,步长分别为 1、2、2。将输出的时间边集序列和空间点集序列串联发送到全连接层将,然后通过全局平均池化层将特征转化为标量。最后使用 Softmax 激活函数进行分类,得到手势识别结果。

2 实验与分析

2.1 实验准备

实验环境为 Windows10 操作系统;CPU 为 Intel(R) Core(TM) i7-9750H;GPU 为 GeForce GTX 1650。设计网络使用 PyTorch 深度学习框架实现。

2.2 实验结果与分析

本文使用 CSL 和手 DEVISIGN-L 手势数据集进行实验。CSL 数据集视频长度约为 2 s,帧宽度为 1 280,帧高度为 720,帧速率为 30 帧/s。从中选取 50 个手势,总共 25 000 个视频,将 90% 的数据集划分为训练集,剩下的 10%作为测试集。DEVISIGN-L 数据集共 24 000 个视频,包含 2 000 个手势样本,将每个视频的帧数调整为 260,其他实验细节与 CSL-500 数据集相同。

实验的初始学习速率设置为 0.1,batchsize 设置为 64,epoch 设置为 10 000,权重衰减设置为 0.000 1。使用 Top-1 与 Top-5 作为准确率指标评估网络性能,准确率可用如下公式计算:

$$\text{准确率} = \frac{\text{正确识别的样本数量}}{\text{测试集中的样本总数}} \tag{4}$$

为了验证本文设计的双流网络结构和改进残差结构的有效性,在 CSL 数据集进行消融实验,以基础 GCN 网络为基线分别对比双流 GCN 网络、GCN+2D-ResNet 网络和 SCR-GAN 网络,实验结果如表 1 所示。可以看出双流 GCN 网络结构相比普通 GCN 网络识别效果更好,识别率提高 8%。在基础 GCN 网络加入改进残差结构后准确率也有不小的上升,提高了 7.3%。因此本文对网络的改进是有效的,两种方式结合大大提高了网络的识别准确率。

表 1 网络改进有效性验证 %

网络方法	Top-1	Top-5
GCN	82.6	89.1
双流 GCN	90.6	94.5
GCN+2D-ResNet	89.9	95.7
SCR-GAN	95.2	98.6

为了进一步验证所提出的 SCR-GAN 网络在手势识别方面的有效性,将该方法与其他主流的方法在 CSL 和 DEVISIGN-L 两个数据集上进行比较,对比实验结果如表 2、3 所示。从表中数据可以看出,在手势识别中,与基于 CNN 网络的方法和基于 RNN 网络的方法相比,基于 GCN 网络的方法识别准确率明显更高,SCR-GAN 网络在 CSL 和 DEVISIGN-L 两个数据集上的准确率分别达到了到达了 96.2% 和 69.3%,识别精度最高,识别效果最好。基于残差的 DSTA-Net 和基于双流结构 2s-AGCN 网络识别准确率都比较高,分别达到了 93.2% 和 95.6%,仅次于本文网络。实验结果证明,残差结构和双流结构相结合,可以增强图像序列提取能力,使整个网络框架具有较好的泛化能力和识别效果。

表 2 CSL 数据集上不同方法的比较 %

方法	Top-1	Top-5
VGG-S[5]	76.3	89.2
Temp Conv+LSTM[6]	73.1	88.6
文献[8]	79.6	90.4
CNN+RNN[9]	80.4	91.5
SLR-Net[11]	92.9	96.3
ST-GCN[12]	91.3	95.3
MD-GCN[13]	92.6	95.7
BLSTM-3D[15]	90.6	94.2
DSTA-Net[16]	93.2	97.8
2s-AGCN[17]	95.6	98.8
Shift-GCN[18]	94.5	97.9
SCR-GAN	96.2	99.2

表 3 DEVISIGN-L 数据集上不同方法的比较 %

方法	Top-1	Top-5
VGG-S[5]	30.8	54.3
Temp Conv+LSTM[6]	23.1	45.6
SLR-Net[11]	55.2	71.4
ST-GCN[12]	58.6	75.9
MD-GCN[13]	60.5	79.2
2s-AGCN[17]	68.2	85.2
SCR-GAN	69.3	88.7

3 结 论

本文提出了一种结合改进残差结构的双流图卷积网络来识别手势。将构建的空间图和时序图分别作为空间流和时间流通道的输入,加快网络训练速度,嵌入改进残差结构增加网络深度,避免了梯度消失等问题,将两通道输出的序列串联转换,输入到输入 Softmax 分类器得到最终结果,识别效果良好。在以后的工作中,将针对手势部分遮挡等复杂条件下如何进行精确识别深入研究,不断优化网络的识别性能。

参考文献

[1] TAN Y S, LIM K M, LEE C P. Hand gesture recognition via enhanced densely connected

convolutional neural network-ScienceDirect[J]. Expert Systems with Applications, 2021,175:114797.
[2] HLAB C, MNB C. Hand gesture recognition method based on HOG-LBP features for mobile devices[J]. Procedia Computer Science, 2018, 126:254-263.
[3] 王万良，杨经纬，蒋一波. 基于运动传感器的手势识别[J]. 传感技术学报，2011，24(12):1723-1727.
[4] ZHANG X, CHEN X, LI Y, et al. A framework for hand gesture recognition based on accelerometer and EMG sensors[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2011, 41(6):1064-1076.
[5] CUI R, LIU H, ZHANG C. A deep neural framework for continuous sign language recognition by iterative training [J]. IEEE Transactions on Multimedia, 2019, 21(7):1880-1891.
[6] PIGOU L, ARON V D O, DIELEMAN S, et al. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video [J]. International Journal of Computer Vision, 2018: 430-439.
[7] 赵金龙，陈春雨，于德海，等. 基于 3D 卷积神经网络的手语动作识别[J]. 通信技术，2021，54(2):327-333.
[8] HUANG S, MAO C, TAO J, et al. A novel chinese sign language recognition method based on keyframe-centered clips[J]. IEEE Signal Processing Letters, 2018, 25(3):442-446.
[9] MASOOD S, SRIVASTAVA A, THUWAL H C, et al. Real-time sign language gesture (word) recognition from video sequences using CNN and RNN[M]. Intelligent Engineering Informatics, 2018:623-632.
[10] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. 5th International Conference on Learning Representations, 2016.
[11] MENG L, LI R. An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network[J]. Sensors, 2021, 21(4):1120.
[12] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[J]. Arxiv Preprint, 2018, Arxiv:1801.07455.
[13] XIONG X, WU H, MIN W, et al. Traffic police gesture recognition based on gesture skeleton extractor and multichannel dilated graph convolution network [J]. Electronics, 2021, 10(5):551.
[14] 龚捷，罗聪，罗琴. 基于注意力机制和残差网络的动作识别模型[J]. 电子测量技术，2021，44(14):111-116.
[15] LIAO Y, XIONG P, MIN W, et al. Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks [J]. IEEE Access, 2019:38044-38054.
[16] SHI L, ZHANG Y, CHENG J, et al. Decoupled spatial-temporal attention network for skeleton-based action recognition[J]. Computer Vision-ACCV 2020, 2020:38-53.
[17] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[J]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 12026-12035.
[18] CHENG K, ZHANG Y, HE X, et al. Skeleton-based action recognition with shift graph convolutional network [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020:180-189.

作者简介

程换新，工学博士，教授，主要研究方向为控制科学与工程、人工智能、深度学习、计算机视觉、图像识别等。
E-mail:1755302185@qq.com

成凯，硕士生，主要研究方向为人工智能、神经网络、图像识别等。
E-mail:969206242@qq.com

程力，工学博士，研究员，主要研究方向为人工智能、大数据分析、互联网网络安全等。
E-mail:1659103953@qq.com

蒋泽芹，硕士生，主要研究方向为人工智能、神经网络、图像识别等。
E-mail:17854273630@163.com