

基于改进 3D 卷积网络的人体动作识别

高海玲¹, 王晓东^{1*}, 章联军¹, 赵仲豪¹, 金建国²

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315211; 2. 浙江德塔森特数据技术有限公司, 浙江 宁波 315048)

摘要: 为解决现有多数视频人体动作识别 3D 卷积方法无法区分信息中各维度的重要和非重要特征问题, 提出了通过门控循环单元(Gated Recurrent Unit, GRU)和空间注意力增强模块构建时空特征处理网络的方法, 基于多级特征融合和多组通道注意力特征选择构建网络, 改进基础网络模型 ResNet3D 对视频人体动作识别中的网络模型. 改进后模型在 2 个公开数据集 UCF101 和 HMDB51 上的准确率分别为 96.42% 和 71.08%, 与 C3D、Two-stream 等网络模型相比, 具有更高的识别准确率.

关键词: 深度学习; 人体动作识别; 3D 卷积; 注意力机制

中图分类号: TP391.4

文献标志码: A

文章编号: 1001-5132 (2023) 03-0016-06

随着多媒体技术的快速发展, 对视频信息的处理分析方法得到广泛研究. 视频人体动作识别作为一个基本的视频任务, 不仅具有实用价值, 而且能为更深层次的任务服务. 视频人体动作可以应用在动作分析、高级人机交互、视频监控、虚拟现实等方面.

传统视频人体动作识别方法主要是基于局部信息和全局信息的手工特征提取, 其中改进的密集轨迹(improved Dense Trajectories, iDT)被广泛认为是其中效果最好的一种^[1]. 然而, iDT 计算复杂度高, 并产生较高的特征维度, 很难做到实时性. 基于深度学习方法是搭建神经网络提取特征, 相比传统方法表达能力更强. 其中长短期记忆网络和卷积神经网络(Convolutional Neural Network, CNN)相结合的方法能够捕获时间顺序和长期相关性. 但长短期记忆网络在卷积顶层只能获取高层动作, 不能获取关键低层动作, 而且训练耗时. 基于图卷积方法, 可以建模高阶不规则的数据结构, 但是不能全面研究对动作识别具有重要意义的非局部关节, 且无法确保在人体骨骼距离较远的动作中将关节较好地融合. 双流网络是一种有效且简单的方法^[2], 由于光流仅表示相邻帧之间的运动

信息, 因此对时间跨度较大的动作不利于建模, 且光流的计算会增加时间成本.

基于 3D 卷积的人体动作识别研究的目的是解决处理视频帧时忽略帧间运动信息的问题. Ji 等^[3]首次提出了 3D 卷积方法, 将多帧视频进行堆叠, 且使用 3D 卷积核进行卷积, 经过连续多次的卷积得到最终特征. 之后, Tran 等^[4]通过实验尝试寻找最优的网络框架, 提出了三维卷积网络, 并解决了需要预先检测视频中人物位置的问题. 残差网络(2D 网络)的出现相较于先前的卷积网络具有占用内存更小的优势, 现已经构建的基于残差网络中的残差组件 T3D(Temporal 3D ConvNet)^[5]、3D 的 ResNeXt(Suggesting the Next Dimension)^[6]、I3D (Inflated 3D ConvNet)^[7]、改进的 C3D(Convolutional 3D Neural Network)^[8]以及改进的 ResNet3D(3D Residual Convolutional Neural Network)^[9]等模型. 之后, Zolfaghari 等^[10]提出了 ECO(Efficient Convolutional Network for Online Video Understanding)框架, 解决了上述方法不能较好地提取整个视频全局信息的问题. 再之后, Feichtenhofer 等^[11]提出了一个通过低帧率和高帧率分别提取视频帧捕获空间语义和运动语义的方法. 与此同时, 也有一些

收稿日期: 2022-10-14.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 浙江省自然科学基金(LY20F010005); 宁波市“科技创新 2025”重大专项(2022T005).

第一作者: 高海玲(1998-), 女, 甘肃白银人, 在读硕士研究生, 主要研究方向: 视频信息处理. E-mail: 1450363642@qq.com

*通信作者: 王晓东(1970-), 男, 浙江上虞人, 教授, 主要研究方向: 多媒体信号处理. E-mail: wxd@nbu.edu.cn

基于差分信息构建网络的方法被提出. 如 Li 等^[12]采用分组卷积方法, 分别提取了相邻视频帧间的信息及运用多帧视频提取全局信息. Wang 等^[13]提出时序差分网络, 该方法能减少计算量、建模时序高效. 最近, Yi 等^[14]将通道空间注意模块依次添加到中间特征图的每个切片张量中, 形成通道和空间注意力. 张聪聪等^[15]同样引入注意力机制, 使网络专注于待识别对象. Zhu 等^[16]提出一种新的跨层注意和中心引导的注意机制, 将特征与来自多个尺度的上下文知识相结合. 但以上模型均忽略了时序信息对模型的影响.

本文通过门控循环单元(Gated Recurrent Unit, GRU)^[17]和空间注意力机制, 基于 3D 网络建模, 同时基于多级特征融合和多组通道注意力特征选择构建模型, 并将该模型在 2 个公开数据集 UCF101 和 HMDB51 上检验. 结果表明, 本文构建的模型与现有的相关模型相比性能有所提升.

1 模型

1.1 模型的整体结构

一般的 ResNet3D 网络是基于局部滑动窗口无差别地提取特征. 可以利用不同的注意力机制增强不同维度特征的重要信息, 抑制非重要信息. 此外, 视频的时序信息是一个重要特征, 对时序信息的处理至关重要.

图 1 为本文提出的视频人体动作识别的整体网络结构实例, 主要通过 GRU 和空间注意力增强

模块构建时空特征处理网络, 基于多级特征进行融合, 通过多段通道注意力模块构建特征选择网络. 具体步骤为: 给定一个视频集, 将每个视频稀疏采样 k 帧, 再将处理后的视频帧输入所构建的网络中进行训练, 获取重要特征. 首先对数据进行初步的特征提取; 其次基于 GRU 和空间注意力增强模块先后对视频的时间序列和空间维度的特征进行提取, 从时间维度学习时序信息, 从空间维度学习空间特征的重要性分数, 增强具有重要信息的注意力关键区域; 再次将不同层次的特征进行融合输出, 从而捕获视频中人体的全局与局部信息; 最后对融合输出后的特征基于分段通道注意力进行选择, 获得具有重要信息的关键通道.

特征提取网络主要通过 GRU 增强对模型的时序特征, 并且通过空间注意力机制增强模块增强模型的重要位置信息.

基于 GRU 时序处理模型如图 1 中 GRU 时序处理模块, 动作“put up”和“put down”, 因为方向不同, 被视为 2 个不同的动作, 时间方向性信息对视频序列较为重要. 对时间序列的处理已经出现较多的优秀模型, 如 LSTM 和 GRU, 两者都可以用于处理较长的时间序列, 但 GRU 相较 LSTM 具有较少的参数. 在模型的初始阶段使用 GRU 模块处理视频的时间序列较好. 首先通过卷积和非线性操作对输入数据进行初步提取, 经过维度变换后采用 k 个双向单层 GRU 获取各个时刻的上下文语义信息, 同时捕捉视频序列中时序距离较远的

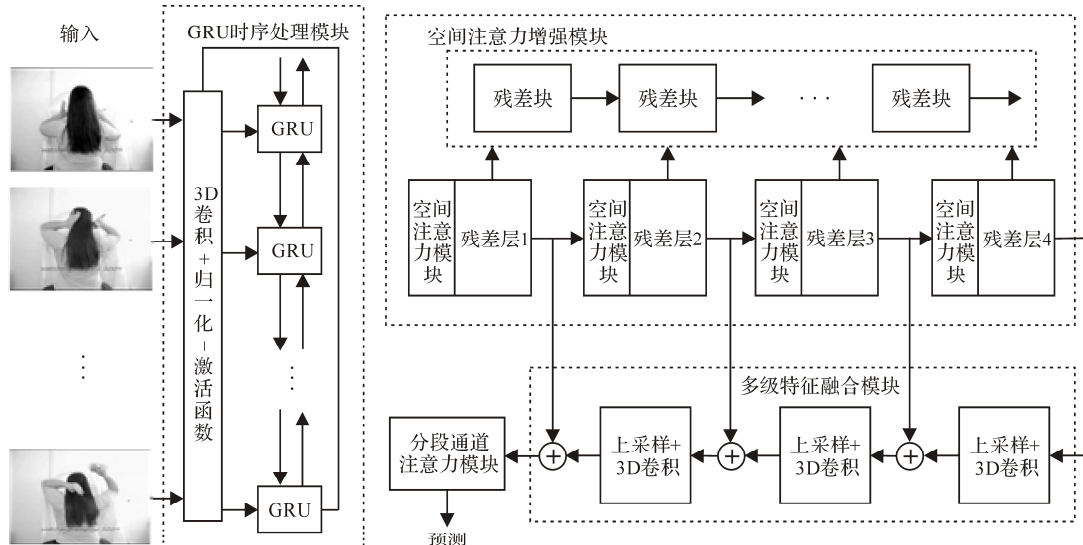


图1 整体网络框图

关系。

基于空间注意力增强特征提取模型如图 2 所示。经过 GRU 时序处理模块处理后,对输出变换获得维度为 (c, t, h, w) 的 X , 其中 c 、 t 、 h 、 w 分别表示输入的通道维度、帧数、每帧视频高和每帧视频宽。首先对 X 的时间维度进行压缩,获得维度为 (c, h, w) 的 X' , 变换维度可以保留关键空间信息;其次经过卷积 1 和卷积 2 两次卷积操作分别对通道数量减少和增加后,提取更多高层信息;再次为每一个通道位置生成对应的权重掩码,并加权输出;最后增强感兴趣的空间区域,得到与 X 相同维度的最终输出 F 。

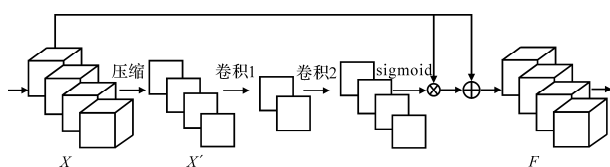


图 2 空间注意力增强模块

空间注意力增强模块嵌入在残差层之前,每一个残差层由多个不同的残差块组成。本文采用 4 个残差层,每个残差层包含的残差块数量分别为 (3, 4, 23, 3)。

1.2 特征融合网络

将特征图 I 进行卷积核为 3×3 的卷积操作后得到感受野为 3×3 特征图 II。将特征图 I 进行两次卷积核为 3×3 卷积操作后得到感受野为 5×5 的特征图 I2, 越高级的特征感受野越大, 越低级特征感受野越小(图 3)。相应地, 感受野越大越关注全局信息, 感受野越小越关注局部信息^[18]。

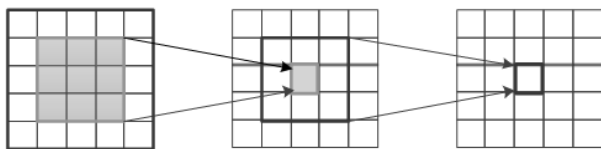


图 3 两次 3×3 卷积图例

为了同时获取全局和局部信息, 本文基于将高级与低级特征相结合思路, 从最后一个残差层的输出开始依次向上融合每两层的输出, 如图 1 中多级特征融合模块所示。残差层 1、残差层 2、残差层 3、残差层 4 的输出维度分别是 $(256, 8, 28, 28)$ 、 $(512, 4, 14, 14)$ 、 $(1\ 024, 2, 7, 7)$ 、 $(2\ 048, 1, 4, 4)$ 。首先对残差层 4 的输出特征进行上采样操作和卷积操

作, 使其维度与残差层 3 的输出维度一致, 然后将两者特征融合; 其次将特征融合后得到的特征进行上采样和卷积操作, 使其维度与残差层 2 的输出维度一致, 然后将两者特征融合; 最后将第 2 次融合后的特征进行上采样和卷积操作, 使其维度与残差层 1 的输出维度一致, 再将两者特征融合, 融合后特征为该模块最终的输出特征。通过该模块能获得视频中不同级别特征的组合信息。

1.3 特征选择网络

通道注意力机制聚焦通道中的重要信息。基于分段注意力模块进行特征选择, 可有效提取通道的重要信息, 弱化不感兴趣的区域, 还可以减少参数量, 提升模型的运行速率。操作如图 4 所示。

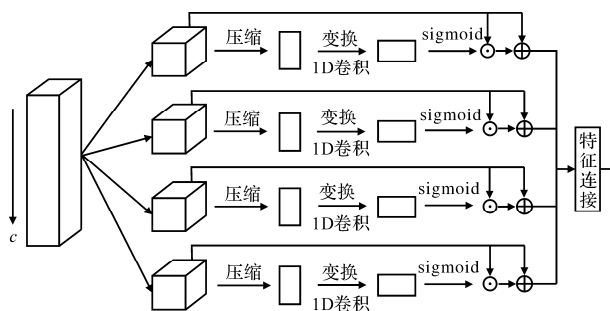


图 4 分段通道注意力增强模块

首先将具有 c 个通道的特征分为 4 段, 使每个通道的数量为 $c/4$; 其次对每一段特征进行通道注意力增强操作, 即将除通道外的时空特征进行池化操作, 使其维度变为 1, 池化后的特征包含时空维度的全局信息, 经过维度变换后对通道维度进行一维卷积操作, 保留通道关键信息并提取更多的高层信息, 经过 sigmoid 函数为每个通道位置生成对应的权重掩码并加权输出, 增强该 $c/4$ 段感兴趣的通道区域; 最后沿通道维度将 4 段的输出特征连接输出。最终分类预测输出是将选择后的特征进行池化和全连接操作后得到。

1.4 损失函数

交叉熵损失函数可以评估当前训练得到的概率分布和真实分布的差异大小。在模型的训练过程中, 使用多分类任务中常用的交叉熵损失函数 L_{cross} , 其公式为:

$$L_{\text{cross}} = \frac{1}{N} \sum_{i=1}^N \sum_{c=1}^M y_{ic} \log p_{ic}, \quad (1)$$

式中: N 表示训练样本的数量; M 表示分类类别的

数量; y_{ic} 表示样本 i 的标签; p_{ic} 表示样本 i 属于第 c 类的预测概率.

交叉熵损失函数主要取决于每个样本对应其分类正确类别的概率, 若该样本分类正确概率较高, 则赋予该样本较小损失; 反之, 则赋予该样本较大损失. 对一般模型而言, 随着训练次数的增加, 损失会一直降低, 直至收敛. 训练前期损失降低速度较快, 训练后期损失降低速度较慢.

2 实验

2.1 实验设置

首先对数据集进行处理, 划分为训练集、验证集和测试集. 对数据进行稀疏采样, 从视频中提取 16 帧. 经过数据增强(水平翻转、随机抖动、裁剪等操作)后, 将处理好的数据输入构建的网络中训练. 训练过程使用 ResNet3D 预训练模型对网络初始化. 训练参数设置为: epoch 为 15, 批处理大小为 16, 学习率为 10^{-3} , 12 范数的权重设置为 10^{-2} , 输入视频样本大小为 $16 \times 3 \times 112 \times 112$. 所有实验中的网络均在一个 16G NVIDIA Tesla T4 上使用 pytorch 分布式机器学习系统进行端到端训练.

2.2 数据集

为了验证模型的有效性, 在主流的公开数据集 UCF101^[19]和 HMDB51^[20]上评测模型.

UCF101 数据集共有 13 320 个视频, 101 个类别. 主要包括人与物体互动、身体动作、人与人之间的互动、演奏乐器、体育运动等(图 5(a)).

HMDB51 数据集共有 6 766 个视频, 51 个类别. 主要包括一般面部动作、用物体操纵的面部动作、一般身体动作、身体运动与对象互动、人与人互动的身体动作等(图 5(b)).

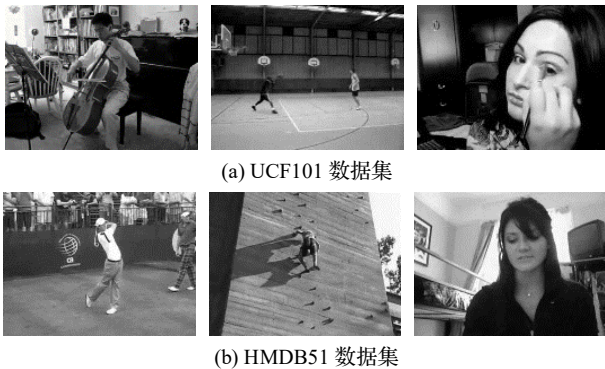


图5 UCF101 和 HMDB51 数据集图例

2.3 实验结果

UCF101 数据集在训练和验证过程中的损失变化和准确率变化如图 6 所示. 从图 6 可见, 在训练集和验证集上的损失变化均随迭代次数的增加而下降, 最终在 0.3 左右处收敛; 在训练集和验证集上准确率变化均随迭代次数的增加而增加, 最终在 0.96 左右处收敛.

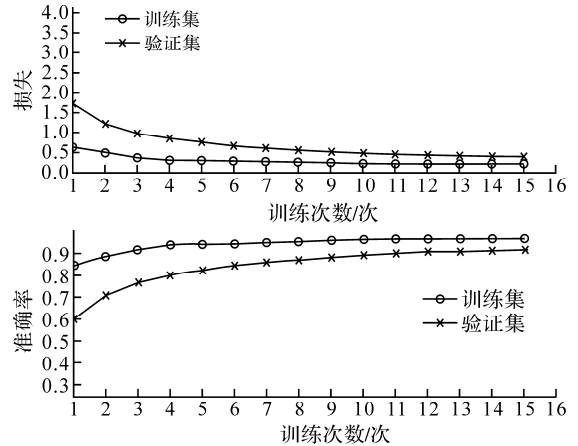


图6 UCF101 数据集上损失和准确率变化曲线

HMDB51 数据集在训练和验证过程中的损失变化和准确率变化如图 7 所示. 从图 7 可见, 在训练集和验证集上的损失变化均随迭代次数的增加而下降, 最终在 1.0 左右处收敛; 在训练集和验证集上的准确率变化均随迭代次数的增加而增加, 最终在 0.71 左右处收敛. HMDB51 数据集相比 UCF101 数据集视频数量少, 背景相对复杂. 因此, 在 HMDB51 数据集上损失较高、识别准确率较低.

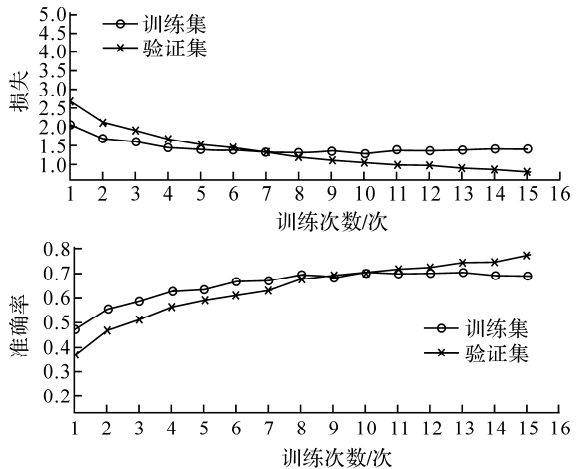


图7 HMDB51 数据集上损失和准确率变化曲线

2.4 网络可视化

网络在 UCF101 和 HMDB51 数据集上的可视

化结果如图8所示. 图8中第1列是从测试样本中随机选取的原图, 第2列是将数据输入网络后获得的特征对应的热力图, 第3列是将热力图乘以权重系数0.5与原图相叠加生成的图像. 从热力图可观察到所构建网络学习的重要区域. 图8(a)聚焦的重点部位是图中人体周围位置, 图8(b)聚焦的重点部位是图中2人之间交互位置, 以此推断网络关注视频数据的空间信息和运动信息.

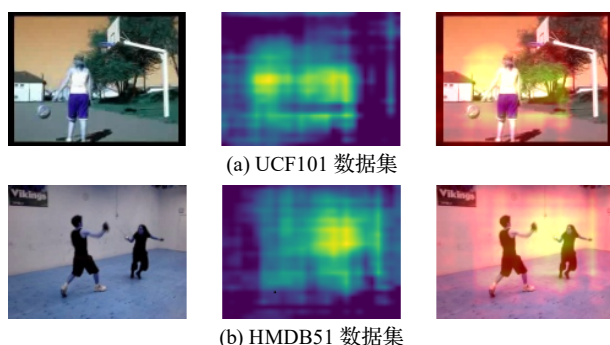


图8 UCF101和HMDB51数据集上网络的可视化结果

2.5 与现有模型性能的比较

表1为不同模型在UCF101和HMDB51数据集上准确率的比较结果.

表1 UCF101和HMDB51数据集上模型准确率比较 %

模型	UCF101 准确率	HMDB51 准确率
iDT ^[1]	86.40	57.20
C3D ^[3]	85.20	—
Two-stream ^[2]	88.00	59.20
3DRseNet50-CS ^[14]	89.90	61.70
T3D ^[5]	93.20	63.50
ResNeXt-101 ^[6]	94.50	70.20
improved DenseNet ^[15]	94.52	69.60
本文方法	96.42	71.08

注: “—”表示对应方法文献中未提供数据.

从表1可见, 本文提出的方法在2个数据集测试中结果均为最优. iDT作为传统人体动作识别的最好方法, 不如表1所列多数基于深度学习的方法. 在UCF101数据集上本文方法比C3D、Two-stream、3DRseNet50-CS、T3D、ResNeXt-101、improved DenseNet方法的准确率分别提升了11.22%、8.42%、6.52%、3.22%、1.92%和1.90%. 在HMDB51数据集上本文方法比iDT、Two-stream、3DRseNet50-CS、T3D、ResNeXt-101、improved DenseNet的准确率分别提升了13.88%、11.88%、9.38%、7.58%、

0.88%和1.48%. 实验结果表明, 本文改进的网络模型在UCF101和HMDB51数据集上均有较好的识别效果.

3 结语

本文提出了一种改进的视频人体动作识别模型, 该模型通过GRU和空间注意力增强模块, 获取视频中的时序信息和空间信息. 在特征融合模块中, 利用多个层级融合获取更多全局重要信息. 在特征选择模块中, 利用分段通道注意力模块获得通道重要信息. 该模型在UCF101和HMDB51数据集上与现有的视频人体动作识别模型相比, 准确率高且有效. 后续将基于光流法提取运动特征作为网络的输入, 构建双流网络模型, 以进一步提高网络识别的准确率.

参考文献:

- [1] Wang H, Schmid C. Action recognition with improved trajectories[C]//2013 IEEE International Conference on Computer Vision, Sydney, Australia, 2014:3551-3558.
- [2] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[EB/OL]. [2022-08-14]. <https://arxiv.org/abs/1406.2199>.
- [3] 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.
- [4] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2016:4489-4497.
- [5] Diba A L, Fayyaz M, Sharma V, et al. Temporal 3D ConvNets: New architecture and transfer learning for video classification[EB/OL]. [2022-08-14]. <https://arxiv.org/abs/1711.08200>.
- [6] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6546-6555.
- [7] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017:4724-4733.
- [8] 张瑞, 李其申, 储珺. 基于3D卷积神经网络的人体动

- 作识别算法[J]. 计算机工程, 2019, 45(1):259-263.
- [9] 刘悦, 张雷, 辛山, 等. 融入时空注意力机制的深度学习网络视频动作分类[J]. 中国科技论文, 2022, 17(3): 281-287.
- [10] Zolfaghari M, Singh K, Brox T. ECO: Efficient convolutional network for online video understanding [C]//European Conference on Computer Vision, Cham: Springer, 2018:713-730.
- [11] Feichtenhofer C, Fan H Q, Malik J, et al. SlowFast networks for video recognition[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2020:6201-6210.
- [12] Li Y, Ji B, Shi X T, et al. TEA: Temporal excitation and aggregation for action recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020:906-915.
- [13] Wang L M, Tong Z, Ji B, et al. TDN: Temporal difference networks for efficient action recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2021:1895-1904.
- [14] Yi Z W, Sun Z H, Feng J C, et al. 3D residual networks with channel-spatial attention module for action recognition[C]//2020 Chinese Automation Congress (CAC), Shanghai, China, 2021:5171-5174.
- [15] 张聪聪, 何宁, 孙琪翔, 等. 基于注意力机制的 3D DenseNet 人体动作识别方法[J]. 计算机工程, 2021, 47(11):313-320.
- [16] Zhu L C, Fan H H, Luo Y W, et al. Temporal cross-layer correlation mining for action recognition[J]. IEEE Transactions on Multimedia, 2022, 24:668-676.
- [17] Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. [2022-08-14]. <https://arxiv.org/abs/1412.3555>.
- [18] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016:2818-2826.
- [19] Soomro K, Amir R Z, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[EB/OL]. [2022-08-14]. <http://export.arxiv.org/pdf/1212.0402>.
- [20] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition[C]//Proceedings of 2011 IEEE International Conference on Computer Vision, USA: IEEE Press, 2011:2556-2563.

Human motion recognition based on improved 3D convolution network

GAO Hailing¹, WANG Xiaodong^{1*}, ZHANG Lianjun¹, ZHAO Shenhao¹, JIN Jianguo²

(1.Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China;

2.Zhejiang DTCT Co., Ltd., Ningbo 315048, China)

Abstract: Video human motion recognition research has great potential for applications, but the modeling quality is greatly affected by movement types, environmental differences and other factors. Most 3D convolution methods for video human motion recognition cannot distinguish between important and non-important features in each dimension given the needed information. To tackle this problem the GRU gating unit and spatial attention enhancement module are used to build a spatio-temporal feature processing network, and the network is built based on multi-level feature fusion and multi-channel attention feature selection. Based on the basic network model ResNet3D, the network model in video human motion recognition is improved. The model achieves 96.42% and 71.08% recognition accuracy on two public datasets UCF101 and HMDB51, respectively, with satisfactory recognition performance. Compared with C3D, two-stream and other generic network models, the proposed model shows higher recognition accuracy, which indicates the effectiveness of the proposed model.

Key words: deep learning; human motion recognition; 3D convolution; attention mechanism

(责任编辑 史小丽)