DOI: 10.13382/j.jemi.B2104768

融合注意力的轻量级行为识别网络研究*

张海超[1] 张 闯[1,2]
(1.南京信息工程大学电子与信息工程学院 南京 210044;2.江苏省气象探测与信息处理重点实验室 南京 210044)

摘 要:针对传统的三维卷积神经网络存在参数量多、信息冗余和时序信息提取不充分 3 个问题,提出了一种融合注意力的轻量级行为识别网络。首先,为轻量化网络参数和融合短中长时序信息,提出了高效残差块来替代两个级联的 3×3×3 卷积;其次,对通道注意力进行拓展,提出了时间注意力机制,并将两者嵌入在网络中抑制冗余信息对识别结果的影响;最后,在 UCF101 数据集上进行实验验证该网络的有效性。结果表明,提出的行为识别网络计算成本为 8.9 GFlops,参数量为 18.0 M,识别准确率为 94.8%,与其他行为识别方法相比,以低成本的计算量实现了较高的识别准确率。

关键词: 3D 卷积神经网络;行为识别;注意力机制;轻量化
中图分类号: TP319.4 **文献标识码**: A **国家标准学科分类代码**: 520.2060

Research on lightweight action recognition network integrating attention

Zhang Haichao[1] Zhang Chuang[1,2]
(1. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing 210044, China)

Abstract: A lightweight action recognition network with fused attention is proposed to deal with the three problems of the traditional 3D convolutional neural network: large number of parameters, information redundancy and insufficient extraction of temporal information. First, in order to lighten the network parameters and fuse short-medium-long temporal information, an efficient residual block is developed to replace two cascaded 3×3×3 convolutions; second, by extending the channel attention mechanism, a temporal attention mechanism is derived, and both of the two mechanisms are integrated into the proposed network to suppress the influence of redundant information on recognition results; finally, experiments are conducted on the UCF101 dataset to verify the effectiveness of the network. The results show that the proposed action recognition network has a computational cost of 8.9 GFlops, a parameter amount of 18.0 M, and a recognition accuracy rate of 94.8%, which reveals a high recognition accuracy with a low cost computation in comparison with other behavior recognition networks.

Keywords: 3D convolutional neural network; action recognition; attention mechanism; light weight

0 引 言

视频行为识别是视频分析领域的代表性任务之一,在视频检索、人机交互、游戏等方面有着广泛的应用,引发了国内外众多学者的关注和研究。但由于视频的拍摄角度、场景、光照和尺度等方面复杂多样,以及远距离时间信息建模难、计算量大,基于视频的行为识别目前还是非常有挑战的研究课题[1-3]。

传统的卷积神经网络[4](convolutional neural networks,CNN)无法提取到相邻帧之间的时序特征,在较长一段时间内,基于 CNN 的视频行为识别方法无法超越人工提取特征的方法[5]。随着研究者对时序信息的关注,Tran 等[6]提出了一种三维卷积神经网络

收稿日期: 2021-09-28 Received Date: 2021-09-28
*基金项目:国家自然科学基金(61906097)、江苏省高校优势学科项目资助

(convolutional 3D, C3D),将二维卷积核增加一个时间维度拓展到三维来提取视频的时空特征,准确率大幅提升且优于人工提取特征的方法。然而相较于2D卷积,3D卷积的三维卷积核使得网络参数量和计算成本呈指数增长,导致网络训练相当耗时且难以满足实际应用中低延时的要求。为此,Xie 等[7]发现在具有高级特征的深层网络里,3D卷积核建模时间特征更加有效,提出将C3D网络中浅层部分的3D卷积替换成2D来降低网络的参数量,但加入了提取相当耗时的光流[8],实用性大打折扣;Tran 等[9]提出由 $1\times d\times d$ 和 $t\times 1\times 1$ 级联的伪3D卷积块替代3D卷积来提取视频序列的时空信息,有效降低了网络参数量和计算成本;张小俊等[10]提出了一种将3D卷积核拆分为由1×3×3的空间流和3×1×1的时间流组成的双流网络[11],通过卷积过程中时间流和空间流特征信息的交互,在减少网络参数的同时,提升了识别准确率。

上述研究者提出的网络只能提取到局部短距离的时序信息,忽略了中长时序信息和信息冗余对网络的影响。Diba 等[12]提出在两个卷积层之间嵌入由不同尺度的 $t\times 1\times 1$ 卷积组成的时间过渡层 temporal transition layer, TTL),以此来提取视频序列的短中长时序信息;Zhu 等[13]通过一组经特殊设计的算子和不受监督的损耗函数,提出了一个全卷积 MotionNet 网络来提取视频帧中与光流相似的运动信息,再与2DCNN提取的空间信息相融合预测行为类别,实现了比双流网络更好的性能。但Diba和Zhu都忽略了卷积过程中存在的信息冗余问题,因此限制了识别准确率的再提高。Liu 等[14]提出了一个由运动增强模块(motion enhanced module, MEM)和时间交互模块(temporal interaction module, TIM)组成的网络结构来代替残差网络中的瓶颈结构[15],其中MEM以相邻帧之间的特征差异为出发点,利用注意力模型来抑制冗余信息的干扰,TIM通过3×1×1卷积拟合时序间的特征信息,该结构有效地解决了信息冗余的问题,但TIM只能拟合短距离时序,忽视了中长时序的拟合。

针对上述视频行为识别方法不能全面有效地解决网络存在的参数量多、时序信息提取不充分和信息冗余3个问题,本文提出了一种融合注意力的轻量级3D卷积神经网络。首先,该网络选择端到端的C3D网络为基础框架;然后,为轻量化网络参数和建模短中长时序信息,本文提出了高效残差块(efficient residual block, ERB)替代两个级联的3×3×3卷积;最后,为抑制网络冗余信息对识别结果的影响,本文先引入通道注意力[16]建模通道之间的相关性进而定位关键特征,再将通道注意力以时间维度展开,拓展为时间注意力,加大对关键帧的关注。为验证ERB、通道注意力和时间注意力对网络的优化作用,本文进行了相关的消融实验。结果表明,3种策略均对网络的识别精度有明显提升,而且将三者联合使用可在减少参数量的前提下保持较高的行为识别精度。

1　方　法

1.1　C3D 网络架构

C3D作为3DCNN[17]中一个经典的网络模型,由于其简洁、紧凑、易于训练和使用的特点,被广泛应用到行为识别、视频相似度分析、动态表情识别等领域。网络结构如图1所示,有8个卷积层、5个池化层、两个全连接层和1个输出的softmax层。其中卷积层使用尺寸为3×3×3的3D卷积核提取视频序列的时空特征,克服了2D卷积只能在空间上学习特征的局限;池化层使用3D池化核为2×2×2的最大池化压缩特征信息,去除冗余信息。

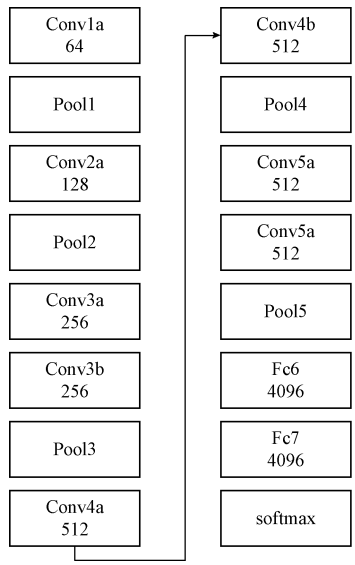


图1　C3D网络结构

Fig. 1　C3D network structure

1.2　高效残差块

为了减少网络参数数量及融合短中长时序信息,本文受瓶颈结构[15]的启发,提出了高效残差块替代两个级联的3×3×3卷积(3×3×3表示该层卷积核尺寸),具体的实现过程为:首先考虑到3×3×3卷积层由于输入和输出维度过大导致参数量和计算成本激增且只能拟合相距为3的时序信息,本文将一个3×3×3卷积层替换为一个瓶颈式卷积块,如图2所示。

图2中 m 表示上一层输出的通道维度,n 表示该层输出的通道维度,卷积块由卷积核分别为3×1×1、3×3×3、3×1×1的3层卷积级联构成。其中第1个3×1×1卷积拟合短距离时序信息并降低输出通道维度为原来的1/4;中间的3×3×3卷积核用于提取时空信息,因其输入和输出维度缩短,参数量大幅减少;最后的3×1×1卷积

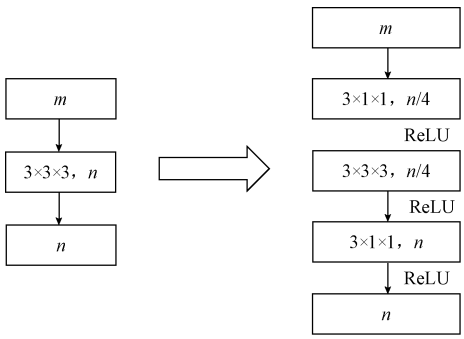


图 2　卷积替换过程

Fig. 2　Convolution replacement process

层提升通道的维度至原大小并再次拟合时序信息。由于所有卷积步长均为 1,替换后的卷积块可拟合相距为 7 的时序信息,卷积块沿时间方向上拟合信息的过程如图 3 所示。

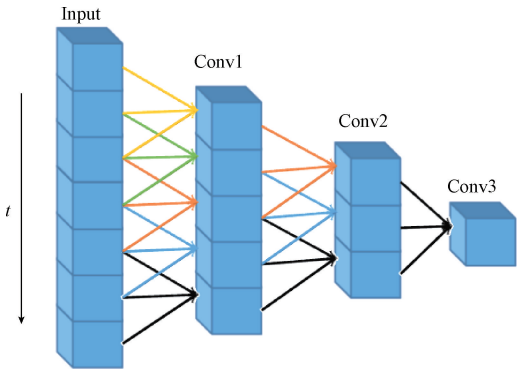


图 3　卷积块沿时间方向上的信息拟合过程

Fig. 3　Information fitting process of convolution block along the time direction

图 3 中输入是相距为 7 的时序信息(每个小正方体表示一个时刻的特征信息),由于卷积核大小为 3、步长为 1,卷积层输出的每个正方体都拟合了前一层对应 3 个时刻的特征信息。以此类推,经过 3 个卷积层后,Conv3 输出的每个正方体可间接拟合输入 7 个时刻的特征信息。

在参数量和计算量方面,根据当前层参数量 Parameters 计算公式:

$$Parameters = k_t \times k_w \times k_h \times c_i \times c_0 + c_0 \quad (1)$$

当前层计算量 FLOPs 计算公式:

$$FLOPs = 2 \times k_t \times k_w \times k_h \times t \times w \times h \times c_i \times c_0 \quad (2)$$

式中:c_i 为输入的特征图个数,c_0 为输出的特征图个数,k_t、k_w、k_h 为卷积核在时间、宽、高 3 个维度的大小,t、w、h 为输入特征图的时间长度、宽和高。将图 2 中替换前后的卷积和卷积块代入式(1)、(2),可得替换后的卷积块参数量和计算量降低了约 8.5 倍。

考虑上述卷积块只能提取相距为 7 的时序信息且加深网络层数易导致网络退化[18],本文受密集连接思想[19]的启发,将两个级联的卷积块密集连接,具体结构如图 4 所示。

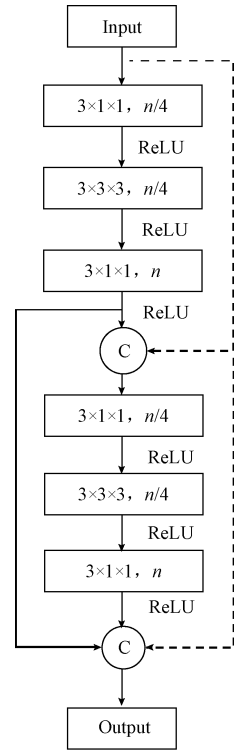


图 4　ERB 结构

Fig. 4　ERB structure

图 4 中 C 表示拼接,实线表示通道数相同可直接拼接,虚线表示通道数不同,需经过 1×1×1 卷积调整通道维度才可拼接。ERB 内部共使用两次拼接操作来融合特征信息,总的来说,ERB 有如下 3 个优点:

1)增加了网络的非线性表达能力。在每层卷积后添加非线性激活(ReLU),更好地拟合了通道、时序间的相关性,使网络学习到更复杂的时空特征。

2)降低参数量、计算量的同时 3×1×1 卷积可拟合短时间信息。

3)融合了输入和两个卷积块输出的短中长时序特征,增加时序信息的多样性,提高网络提取行为特征的鲁棒性。

1.3　通道和时间注意力

源于人类视觉会选择性的关注有用信息,忽略其他可见信息,近年来,模拟人类视觉的注意力机制[20]被广泛应用于神经网络。为了区分各特征图的重要程度,本文引入了通道注意力机制[16]来重新分配特征图之间的权重,加大关键特征图的关注,弱化冗余特征图对识别结果的影响。考虑到输入视频段中不同帧对判断行为类别

起的作用不同,如跳高,表示跳的帧往往比助跑的帧更加关键,本文将通道注意力拓展到时间域,提出了如图5所示的时间注意力机制。

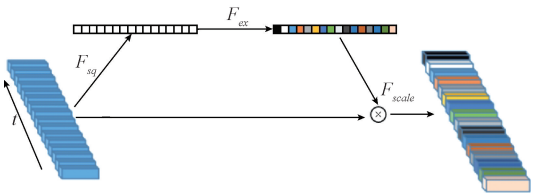


图5　时间注意力机制

Fig. 5　Time attention mechanism

图5中注意力机制的输入是维度为 $t\times h\times w\times c$ 的特征图,输出为维度相等的被赋予不同权值的特征图;F_{sq}、F_{ex}、F_{scale} 分别压缩、激励、权重分配3个函数。其中,压缩是对输入的特征图沿时间方向展开,对每个时刻的特征信息做全局平均实现信息的压缩,表达式为:

$$Z_t = F_{sq}(U_t) = \frac{1}{C\times H\times W}\sum_{i}^{C}\sum_{j}^{H}\sum_{k}^{W}U_t(i,j,k) \quad (3)$$

式中:U_t 代表输入特征图,C 表示特征图的通道数,H、W 表示单张特征图的高和宽。

激励是通过两个全连接层以训练的方式学习每个时刻的权重,表达式为:

$$S_t = F_{ex}(Z_t,W) = \sigma(g(Z_t,W)) = \sigma(W_2\delta(W_1Z_t)) \quad (4)$$

式中:σ、δ 分别代表激活函数Sigmoid和ReLU,W_1、W_2 是两层神经网络的权重,Z_t 为输入特征图在时间维度上的压缩特征。

权重分配是将输入乘以对应时刻的激励权重,从而增强对关键帧的注意,表达式为:

$$X_t = F_{scale}(Z_t,S_t) = Z_tS_t \quad (5)$$

1.4　本文网络模型

本文的网络模型整体框图如图6所示,以C3D网络为基础架构,前两层卷积后嵌入时间注意力机制,3、4、5层中以高效残差块(ERB)替换两个相连的卷积层,考虑到ERB以拼接的方式融合前层网络信息会使网络增宽,本文在第6层加入1×1×1卷积降将网络的维度降低1/2,然后接通道注意力机制,后续每个全连接层的神经元为2 048个,输出层保持不变。

图6中网络的输入为16张连续的,大小为112×112×3的视频帧,卷积核步长均为(1,1,1),且使用padding保持输出特征图尺度不变。池化层中,除了第1层池化核采用步长为(1,2,2)保留时间信息外,其余池化核均采用步长为(2,2,2)的最大池化,每经过一个池化层,特征图尺度压缩为原来1/8。

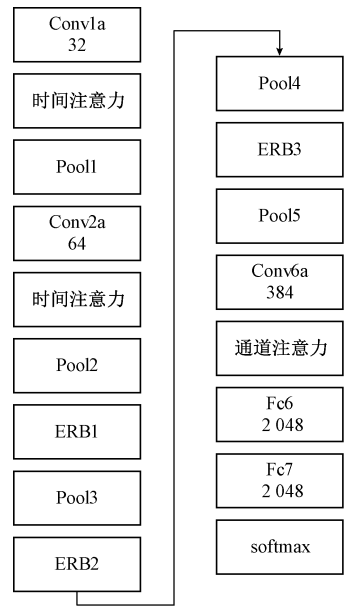


图6　网络结构框图

Fig. 6　Network structure block diagram

2　实验与分析

2.1　数据集

本文选用UCF101行为识别数据集[21]进行实验验证。UCF101数据集含有101类行为动作,图7为该数据集部分动作示例,每类动作被分为25组,每组包含一个动作的4~7个视频,总计13 320个视频。动作涉及人与物体交互、单纯的肢体动作、人与人交互、演奏乐器、体育运动5个方面。

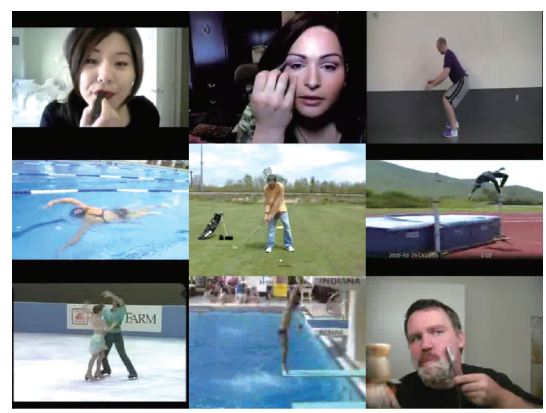


图7　UCF101数据集部分示例

Fig. 7　Some examples of UCF101 data set

对数据集的预处理,首先以FPS为10将视频切分为一张张静态的图片保存在本地,然后以4∶1的比例,随机将切分好的静态数据集划分为训练集和测试集,最后为了防止训练过程中网络出现过拟合的现象,采用水平

翻转和亮度调整两种策略对训练集的数据进行增强。

2.2　实验设置

1)实验环境

本文在 Intel(R) Core(TM) I5-9400F, NVIDIA GeForce GTX 1600 SUPER(6 GB),2.9 GHz CPU、64 位 Windows10 操作系统上进行实验,采用 Python 语言编程,TensorFlow-Slim 轻量级库搭建神经网络模型。

2)超参数设置

训练阶段,随机初始化网络参数开始训练网络,采用小批量数据(mini-batch)进行训练,batch 大小为 16;反向传播时,使用 Adam 优化算法[22]训练模型参数,学习率使用分段常数衰减策略,前 35 个 epoch 设置为 0.000 1,之后每 15 个 epoch 衰减为原来 1/10;采用 L2 正则化和 Dropout 两种策略防止网络出现过拟合,参数分别设置为 0.000 5 和 0.5。

测试阶段,在测试集中取一段视频随机连续的 16 帧为网络输入,通过前向传播输出 101 个行为分类得分,取得分最高的类别为预测结果,最终为保证测试结果的鲁棒性,取 10 次测试集的平均准确率为最终的评估指标。

2.3　高效残差块参数选择实验

高效残差块对输入数据中长时序信息的获取是通过两个卷积块实现的,因此可调整卷积块中前后两个 $t\times1\times1$ 卷积核尺寸改变卷积块在时间维度上提取的深度。本文以 C3D 网络(用 $t=0$ 表示)为参考对象,比较了 t 分别为 1、3、5 时,高效残差块对识别准确率的影响,结果如表 1 所示。

表 1　高效残差块中 t 取不同值时在 UCF101 上的准确率

Table 1　The accuracy on UCF101 when t takes different values in the high-efficiency residual block

t	准确率/%
0	89.9
1	92.3
3	92.5
5	92.1

从表 1 中可见,加入高效残差块后,无论 t 为何值都比原 C3D 网络的分类精度高,表明本文所提出的高效残差块确实可以优化网络,提升识别准确率。其中当 $t=3$ 时准确率明显优于其他取值,因为本文输入视频长度为 16,$t=1$ 时高效卷积块只能融合距离为 1、3、5 的时序信息,缺少了对更长时序信息的融合;$t=5$ 时恰好相反,融合了距离为 1、11、16 的短长时序信息,忽略了中时序信息的融合。$t=3$ 时,正好可融合距离为 1、7、13 的短中长时序信息,使得模型的精度更高。

2.4　通道和时间注意力的消融实验

考虑到不同注意力机制对网络的影响,本文分别测试了不使用注意力(表 2 中用 A 表示)、仅使用时间注意力(表 2 中用 B 表示)、仅使用通道注意力(表 2 中用 C 表示)、通道和时间注意力都使用(表 2 中用 D 表示)4 种方案的识别准确率,结果如表 2 所示。

表 2　不同注意力机制加入网络后在 UCF101 上的准确率

Table 2　The accuracy of different attention mechanisms on UCF101 after joining the network

方案	准确率/%
A	92.5
B	93.5
C	93.4
D	94.8

由表 2 可见,B、C、D 这 3 种方案都比 A 方案的识别准确率高,这表明注意力机制关注到了关键的信息从而优化了网络。相较于 B 和 C 方案,两者的作用域不同,B 方案更加关注时间维度上的关键帧,C 方案更加关注通道维度上不同特征图的重要程度,两种方案对准确率都有提升,但是差别不大,说明单个注意力机制对网络的作用是有限。因此可将 B、C 方案联合起来(D 方案)让网络同时关注时间和维度方面的关键信息以提高识别准确率。

2.5　与其他行为识别方法比较

经上述所做的对比实验,本文选择 $t=3$ 的高效残差块以及联合使用通道和时间注意力机制为最终的网络模型,为了验证本文所提出方法的优势,将该模型与当下流行的行为识别方法在不同的性能指标上的比较,结果如表 3 所示。

表 3　不同识别方法在 UCF101 数据集上性能对比

Table 3　Performance comparison of different recognition methods on UCF101 data set

方法	是否使用光流	浮点运算量/GFLOPs	参数量/M	准确率/%
IDT[5]	是			85.9
C3D[6]	否	38.5	72.9	82.3
R(2+1D)[9]	否	152.4	33.3	**96.8**
改进 C3D[10]	否	14.4	26.5	90.7
T3D[12]	否	19.8	85.5	93.2
TSN[23]	是	**3.8**	33.6	94.0
Dynamo Net[24]	否		42.1	93.1
文献[25]	是	4.1	25.5	93.8
本文方法	否	8.9	**18.0**	94.8

表 3 中浮点运算量指网络模型由输入到输出执行一次所需要的计算量。具体的计算方式为:先根据式(2)得到每个卷积层和全连接层的计算量,再将每层计算量相加得到整个网络的计算量,其表示网络的计算复杂度,计算量越小,网络的运行速度越快。从表 3 中可以看出,

文献[9]的计算量最大,会使网络的训练周期变长、实时性变差;文献[23,25]都使用到了光流,虽然网络的计算量较低,但光流图的提取相当耗时,严重影响网络的实时性;文献[6,10,12]和本文的方法都是以3D CNN为架构而设计的识别方法,但在计算量上分别是本文的13.3、1.6、2.2倍。综上所述,在运行速度方面,相较于其他识别方法,本文表现出了更好的性能。

对表3整体分析可知,本文提出的方法在不使用光流的前提下,无论是计算量、参数量,还是识别准确率都明显优于大部分网络。其中文献[9]准确率高于本文,但计算量和参数量分别是本文的17.1倍和1.6倍;文献[23,25]中计算量都低于本文,但是两者的参数量分别是本文的1.8倍和1.4倍,准确率也比本文低0.8%和1.0%。总的来说,本文所提出的方法较好的均衡了计算成本、参数量和准确率3个性能指标,在降低参数量、计算成本的同时保证了较高的准确率。

3 结 论

本文针对3D卷积过程中存在的参数量多、信息冗余和时序信息提取不充分3个问题,提出了一种融合注意力的轻量级行为识别网络。本文的主要贡献为:1)提出了ERB代替两个级联的3×3×3卷积来降低网络的参数量、提取视频序列的短中长信息,实验表明,ERB具有更好的性能。2)将通道注意力拓展为时间注意力,并将两者都嵌入在网络中,实验表明,同时使用通道和时间注意力效果更佳。最后,通过与经典的行为识别方法对比,论证了本文提出的方法具有少的参数量和更高的识别准确率。

参考文献

[1] ZHU Y, LI X Y, LIU C H, et al. A comprehensive study of deep video Action recognition[J]. arXiv preprint arXiv:2012.06567, 2020.

[2] 周育新, 白宏阳, 李伟. 基于关键帧的轻量化识别方法研究[J]. 仪器仪表学报, 2020, 41(7): 196-204.
ZHOU Y X, BAI H Y, LI W. Research on lightweight behavior recognition method based on key frame[J]. Chinese Journal of Scientific Instrument, 2020, 41(7): 196-204.

[3] 王丽君, 刘彦戎, 王丽静. 基于卷积长短时深度神经网络行为识别方法[J]. 电子测量与仪器学报, 2020, 34(9): 160-166.
WANG L J, LIU Y R, WANG L J. A deep neural network behavior recognition method based on long and short convolutions[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(9): 160-166.

[4] LI Z W, LIU F, YANG W J, et al. A survey of convolutional neural networks: Analysis, applications, and prospects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021: 1-21.

[5] WANG H, SCHMID C, CORDELIA. Action recognition with improved trajectories[C]. 2013 IEEE International Conference on Computer Vision, 2013: 3551-3558.

[6] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 4489-4497.

[7] XIE S N, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 318-335.

[8] 邵绪强, 杨艳, 刘艺林. 流体运动估计光流算法研究综述[J]. 中国图象图形学报, 2021, 26(2): 355-367.
SHAO X Q, YANG Y, LIU Y L. Review of optical flow algorithm for fluid motion estimation[J]. Journal of Image and Graphics, 2021, 26(2): 355-367.

[9] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 6450-6459.

[10] 张小俊, 李辰政, 孙凌宇. 基于改进3D卷积神经网络的行为[J]. 计算机集成制造系统, 2019, 25(8): 2000-2006.
ZHANG X J, LI CH ZH, SUN L Y. Behavior recognition based on improved 3D convolutional neural network[J]. Computer Integrated Manufacturing System, 2019, 25(8): 2000-2006.

[11] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. Advances in Neural Information Processing Systems, 2014: 568-576.

[12] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3D convnets: New architecture and transfer learning for video classification[J]. arXiv preprint arXiv: 1711.08200, 2017.

[13] ZHU Y, LAN Z Z, NEWSAM S, et al. Hidden two-stream convolutional networks for action recognition[C]. Asian Conference on Computer Vision, 2018: 363-378.

[14] LIU Z Y, LUO D H, WANG Y B, et al. Teinet: Towards an efficient architecture for video recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11669-11676.

[15] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

[16] HU J, LI S, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[17] 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.

[18] HE K M, SUN J. Convolutional neural networks at constrained time cost[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 5353-5360.

[19] HUANG G, LIU Z, VAN DER MATTEN L, et al. Densely connected convolutional networks[C], 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2261-2269.

[20] CHAUDHARI S, MITHAL V, POLATKAN G, et al. An attentive survey of attention models [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-32.

[21] SOOMRO K, ZAMIR A, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

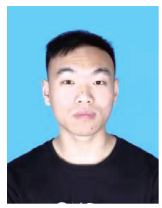
[22] DE S, MUKHERJEE A, ULLAH E. Convergence guarantees for RMSProp and ADAM in non-convex optimization and an empirical comparison to Nesterov acceleration [J]. arXiv preprint arXiv: 1807.06766, 2018.

[23] WANG L, XIONG Y J, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2016: 20-36.

[24] DIBA A, SHARMA V, VAN GOOL L, et al. Dynamonet: Dynamic action and motion network[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 6191-6200.

[25] 张冰冰, 葛疏雨, 王旗龙. 基于多阶信息融合的行为识别方法研究[J]. 自动化学报, 2021, 47(3): 609-619.
ZHANG B B, GE SH Y, WANG Q L. Research on behavior recognition method based on multi-level information fusion[J]. Acta Automatica Sinica, 2021, 47(3): 609-619.

作者简介



张海超,2020 年于华北水利水电大学获得学士学位,现为南京信息工程大学硕士研究生,主要研究方向为深度学习和行为识别。
E-mail:1015661013@qq.com

Zhang Haichao received his B. Sc. degree from North China University of Water Resources and Electric Power in 2020. Now he is a M. Sc. candidate at Nanjing University of Information Science and Technology. His main research interests include deep learning and behavior recognition.

张闯(通信作者)1998 年于河北师范大学获得学士学位,2004 年于燕山大学获得硕士学位,2008 年于南京理工大学获得博士学位,现为南京信息工程大学副教授,主要研究方向为光电信息及视觉信息采集与处理。
E-mail:zhch_76@163.com

Zhang Chuang (Corresponding author) received her B. Sc. degree from Hebei Normal University in 1998, M. Sc. degree from Yanshan University in 2004, and Ph. D. degree from Nanjing University of Science and Technology in 2008. Now she is an associate professor at Nanjing University of Science and Technology. Her main research interests include photoelectric information and visual information acquisition and processing.