DOI:10.19651/j.cnki.emt.2106848

基于注意力机制和残差网络的动作识别模型*

龚 捷 罗 聪 罗 琴
(西南石油大学 计算机科学学院 成都 610599)

摘 要: 深度学习在图像领域取得的突破,使得特征学习方面取得了迅猛的发展。针对视频序列中连续帧具有的时间相关性,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。首先利用残差 3D 卷积网络学习视频序列中连续视频帧之间的时间相关性,即时空特征;之后利用扩展到三维的通道注意力网络对残差 3D 卷积结构学习到的每个特征通道赋予不同的权值;最后将重新标定权重的特征输入分类器得到最终的分类。在 UCF-101 和 HMDB-51 数据集上进行实验,分别取得了 95.8%和 69.7%的准确率。实验结果表明,所提出的模型在视频人体动作识别问题上具有较高的识别准确率。

关键词: 残差网络;三维卷积;注意力机制;深度学习;动作识别
中图分类号: TP391.4 **文献标识码:** A **国家标准学科分类代码:** 520.2

Action recognition model based on attention mechanism and residual network

Gong Jie Luo Cong Luo Qin
(School of Computer Science, Southwest Petroleum University, Chengdu 610599, China)

Abstract: The breakthrough of deep learning in the field of image makes the rapid development of feature learning. Aiming at the temporal correlation of consecutive frames in video sequences, a residual 3D convolutional network model based on attention mechanism is proposed for human action recognition. Firstly, residual 3D convolution network is used to learn the temporal correlation between consecutive video frames in video sequence. Then, each feature channel learned by residual 3D convolution structure is given different weights by using channel attention network which is extended to three-dimensional. Finally, the reweighted features are input into the classifier to get the final classification. Experiments are carried out on UCF-101 and HMDB-51 datasets, and the accuracy is 95.8% and 69.7%, respectively. The experimental results show that the proposed model has high recognition accuracy in video human action recognition.

Keywords: residual network;3D convolution;attention mechanism;deep learning;action recognition

0 引 言

计算机视觉领域几十年来一直致力于视频分析,并提出了不同的问题,如文献[1]提到的人体动作识别、文献[2]提到的异常事件检测以及文献[3]提到的活动理解。通过采用不同的具体解决办法,在这些个别问题上取得了相当大的进展。人体动作识别的主要目的是从一段视频或连续图片序列中分析识别出一个或多个人物的动作。然而,与静态图像的理解相比,在视频领域的网络结构设计和特征学习方面的进展缓慢,部分原因在于视频数据固有的复杂性和更高的维度。

自从文献[4]提出 AlexNet 网络以来,深度学习经过了更加深入的创新研究(如更小的空间滤波器、多尺度卷积、剩余学习和密集连接)。在图像识别领域,文献[5]中所使用的卷积神经网络(CNN)在多个领域被证明是有效的,例如文献[6]提到的图像分类、文献[7]提到的目标检测和文献[8]提到的语义分割。深度学习在图像领域取得的突破,使得特征学习方面也有了迅猛发展,例如 Zhang 等[9]提出的各种预训练卷积神经网络(ConvNet)模型被用于图像特征的提取,这些特征被用于激活整个网络模型的全连接层,这些层在迁移学习任务中表现良好。然而,由于视频是一种具有时序特征的变化数据,任意像素与其相邻像素之间

收稿日期:2021-06-03
* 基金项目:国家自然科学基金(61902328)项目资助

的相似性很大，具有很强的时间相关性与空间相关性。而卷积神经网络通常用于单一、静止的图片，不能有效地提取出视频序列中的这种时空特征。

与识别图像中的人体动作相比，视频序列中的人体动作是由视觉表象组成的三维信号，随着时间的推移而动态演化。为了学习视频序列中的时空特征，有人试图改变二维卷积神经网络或利用其他深层网络模块来编码动作的时间信息。Tran 等[10]提出的三维 CNN 模型(C3D)通过执行三维卷积从空间和时间维度提取特征，从而捕获编码在多个相邻帧中的运动信息。其基础是将已建立的 2D CNN 体系结构直接扩展到三维时空域，训练出一个全新的 3D 卷积神经网络模型。

随着网络结构的加深，3D 卷积神经网络模型能够学习到视频序列中的时空特征也随之增多。但是由于梯度爆炸和梯度消失问题，在测试数据和训练数据中的准确率却有所下降。为了能够在深层网络中得到更好的训练效果，本文采用的网络结构将文献[11]提出的残差网络结构与 C3D 网络结构相结合，同样使用 3D 卷积来学习时空特征，同时将网络的结构设计限制在残差网络框架中。对于视频中的人体动作而言，某些部位的动作相较于其他部位来说更加重要，例如在射箭过程中，人体躯干在整个动作进行时都处于静止不动的状态，我们的关注点应该着重在手部的运动，从而判断动作类别。因此，在对动作特征进行提取之后应该对不同位置的动作特征赋予不同的权重。

本文结合了 3D 卷积和残差网络结构的特点，构建了一种基于注意力机制和残差 3D 卷积网络的人体动作识别模型。该模型首先利用残差 3D 卷积网络学习视频序列中连续视频帧之间的时空特征；之后利用扩展到三维结构的注意力网络对每个特征通道赋予不同的权值；最后将重新标定权重的特征输入分类器得到最终的分类。构建出本文的网络结构后，在 UCF-101 和 HMDB-51 数据集上进行实验。通过对比其他网络结构在此数据集上的实验结果，可以看出本文所提出的模型在动作识别问题上表现更好。

1　相关工作

视频数据的特征学习方法可以分为两类：基于手工制作的方法和基于深度学习的方法。基于手工制作的特征表示包括 STIPs、SIFT-3D、HOG3D、Cuboids 和 ActionBank。这些手工制作的特征表示使用不同的特征编码方案，如特征直方图或金字塔。在手工制作的表示法中，文献[12]提到的改进的密集轨迹(iDT)被称为当前最先进的手工制作特征，在不同的视频分类问题上具有强大的效果。

基于深度学习的时空特征表示方法有 3 种成功的结构框架：3D CNN 结构、双流 CNN 结构和顶部有时间模型的二维 CNN 结构。3D CNN 结构已被证明在大规模数据集上训练时可以产生很强的动作识别效果。3D CNN 结构的特性也被证明可以很好地推广到其他任务，包括动作检测、视频字幕和手势检测。Simonyan 等[13]提出的双流 CNN 结构，首先利用传统的 CNN 模型从视频序列的光流数据中提取出深层特征以及从连续帧的 RGB 图像中提取出空间特征，在之后的全连接层中将两种特征相融合，从而得到视频序列的时空特征。顶部有时间模型的二维 CNN 结构，在原始二维 CNN 结构的基础上加入了时间特征，如文献[14]提出的长短时深度神经网络，该结构能够提取到视频序列中的连续时间特征，从而提高模型的识别准确率。

Chaudhari 等在文献[15]中提到，注意力机制的基础是利用人类视觉机制，视觉系统中倾向于关注图像中辅助判断的部分信息，并忽略掉不相关的信息，所以使用注意力机制对于视频中复杂特征信息的提取和利用是有效的。Jaderberg 等[16]提出了一种注意力模块，通过对图像进行空间变换来提取图像中的关键信息。Hu[17]提出了通道注意力模块(squeeze-and-excitation networks，SENet)，建模通道之间的相互依赖关系，通过网络的全局损失函数自适应地重新矫正通道之间的特征相应强度。Zhang 等[18]提出了空间自注意力机制，使得网络对感兴趣区域有更多的关注。本文为了更好地提取视频中动作的特征，在残差 3D 卷积网络中嵌入扩展到三维的通道注意力网络以提高其性能。

2　残差 3D 卷积结构

2.1　3D 卷积

与 2D 卷积相比，3D 卷积由于采用了三维卷积计算和三维池化操作，能够更好地对视频序列中的时间信息进行建模。因此，3D 卷积非常适合时空特征的学习。在 3D 卷积中，卷积计算和池化操作是在时空上执行的，在 2D 卷积的基础上添加了时间维度，通过堆叠多个连续的帧组成 1 个立方体，然后在立方体中运用 3D 卷积核，如图 1 所示。

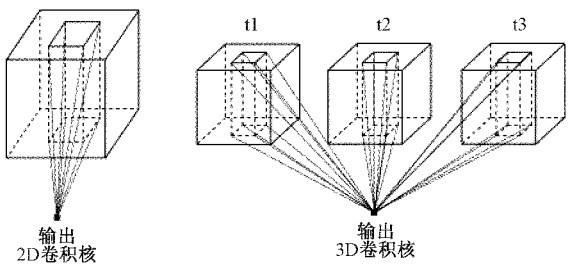

图 1　3D 卷积示意图

当输入视频流数据时，相当于输入了多帧连续的图像数据。对于传统的 2D 卷积来说，是将简单的 2D 卷积核(形如 $i\times i$)运用在每一帧图像上，此时仅仅对空间维度进行了卷积计算，所以会造成精度低、分类不准确的问题。3D 卷积将 2D 卷积的卷积核膨胀为 $i\times i\times i$ 的形式。多个连续帧依次通过卷积层，卷积层中每个特征图都与上一层的多个相邻连续帧相连，从而获取一定的运动信息。整个计算过程如式(1)所示。

$$v_{i,j}^{x,y,t}=\tanh\left(\sum_{k=1}^{m}\sum_{p=1}^{P_i}\sum_{q=1}^{Q_i}\sum_{r=1}^{R_i}w_{i,j,k}^{p,q,r}v_{i\ 1,k}^{x+p,y+q,t+r}+b_{i,j}\right) \tag{1}$$

式中:$v_{i,j}^{x,y,t}$ 表示在第 i 层第 j 个特征图第 t 个通道位于像素点(x,y)处的值;m 表示第$(i-1)$层的特征图个数;P_i、Q_i、R_i 表示第 i 层的 3D 卷积核的空间维度与时间维度大小;$w_{i,j,k}^{p,q,r}$ 表示前一层第 m 个特征图连接的权重大小;$b_{i,j}$ 表示第 i 层第 j 个通道的偏置。

2.2　残差 3D 卷积

针对视频序列数据中连续视频帧之间存在的相关性,利用 3D CNN 学习其中的时空特征。本文使用的 3D CNN 基础结构如图 2 所示,该 3D CNN 结构在学习到视频序列中连续帧之间视觉外观的同时,也可以提取出视频序列中连续帧之间时间的演化,即学习到视频序列的时空特征。

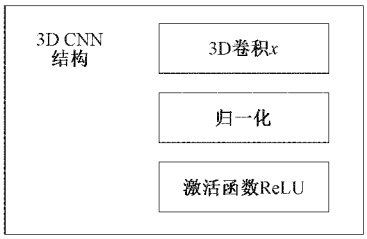

图 2　3D 卷积结构

其中,x 表示卷积核的个数,所有卷积核的大小都是 3×3×3,所有这些卷积层都应用了适当的填充(包括空间和时间)且步长尺寸为 1×1×1,因此从这些卷积层的输入到输出的大小没有变化。

随着网络结构的加深,3D CNN 结构能够学习到的视频序列中的时空特征也随之增多。但是由于梯度爆炸和梯度消失问题,在测试数据和训练数据中的准确率反而有所降低。为了解决该问题,本文将 3D CNN 结构与残差网络结构相结合,构建了一种残差 3D 卷积结构,如图 3 所示。

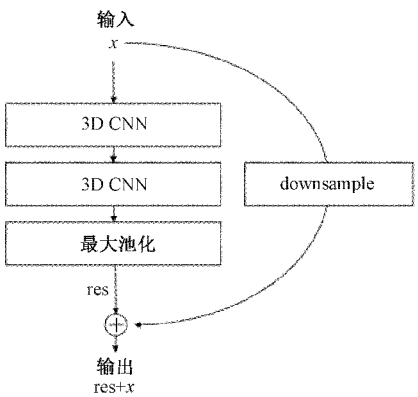

图 3　残差 3D 卷积结构

首先将输入数据通过两个 3D 卷积层以及一个最大池化层,用于学习输入数据的时空特征;同时对输入数据进行 downsample 操作,使得输入数据 x 和学习到的时空特征 res 具有相同的尺寸;最后将两者相加,从而解决梯度消失的问题。

将该结构与注意力机制相结合,为学习得到的时空特征赋予不同的权重从而提升识别准确率,具体的结构在第 3 章中详细介绍。

3　模型结构

3.1　注意力机制

注意力机制可以认为是一种资源分配的机制,对于原本平均分配的资源根据注意力对象的重要程度重新分配资源。在深度神经网络的结构设计中,注意力机制所要分配的资源就是权重。

通道注意力模块从特征通道之间的关系入手,对特征通道间的相互依赖关系进行建模。具体来说,就是通过学习的方式来自动获取每个特征通道的重要程度,然后依照这个重要程度去增强有用的特征并抑制对当前任务用处不大的特征。能够让网络利用全局信息有选择地增强有益特征通道并抑制无用特征通道,从而实现特征通道的自适应校准。具体的操作过程如图 4 所示。

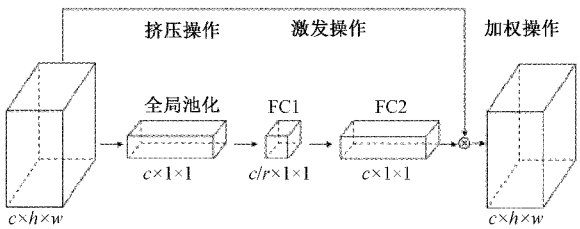

图 4　注意力机制示意图

首先是挤压(Squeeze)操作,采用全局平均池化按照空间维度对特征进行压缩,将每个二维的特征通道变成一个实数,这个实数从某种程度上来说具有全局的感受野,并且输出的维度和输入的特征通道数相匹配。它表征着在特征通道上响应的全局分布,并且使得靠近输入的层也可以获得全局感受野。

其次是激发(Excitation)操作,该操作类似于循环神经网络中的门机制。通过参数学习来为每个特征通道生成权重,并且使用该参数对特征通道间的相关性进行建模。

最后是一个加权(Reweight)操作,将 Excitation 操作的输出权重看作是经过特征选择后每个特征通道的重要性,然后通过乘法逐通道加权到先前的特征上,完成在通道维度上对原始特征的重标定。

3.2　基于注意力机制的残差网络

本文将通道注意力网络直接扩展到三维,对残差 3D 卷积结构学习到的每个特征通道赋予不同的权值,具体结构如图 5 所示。

三维通道注意力网络首先进行 Squeeze 操作,对残差 3D 卷积结构学习到的多尺度特征图进行全局自适应平均池化;之后进行 Excitation 操作,通过两个全连接层来建模

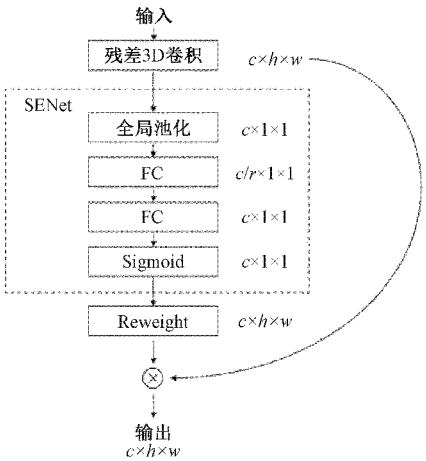

图 5　三维 SENet 结构

通道间的相关性并为每个特征通道生成对应的权重，在扩展到三维的情况下，两个全连接层均采用卷积核大小为 1×1×1 的 3D 卷积。首先通过第 1 个全连接层，将特征维度降低到输入的 1/16；之后再通过第 2 个全连接层使得特征维度升回到原有的大小。利用两个全连接层的好处在于：1）具有更多的非线性，可以更好地拟合通道间复杂的相关性；2）极大地减少了参数量和计算量。然后通过一个 Sigmoid 函数获得归一化后的权重，最后进行 Reweight 操作，得到经过特征选择后的每个特征通道的重要性，然后通过乘法逐通道加权到先前的特征上，完成对原始特征的重标定。

本文设计的网络结构如图 6 所示，该网络结构共有 4 个 SERes 模块，该模块首先利用残差 3D 卷积结构学习输入数据的时空特征，再通过三维 SENet 网络进行 Squeeze-and-Excitation 操作，为原始特征通道分配不同的权重值，从而完成对原始特征的重标定。

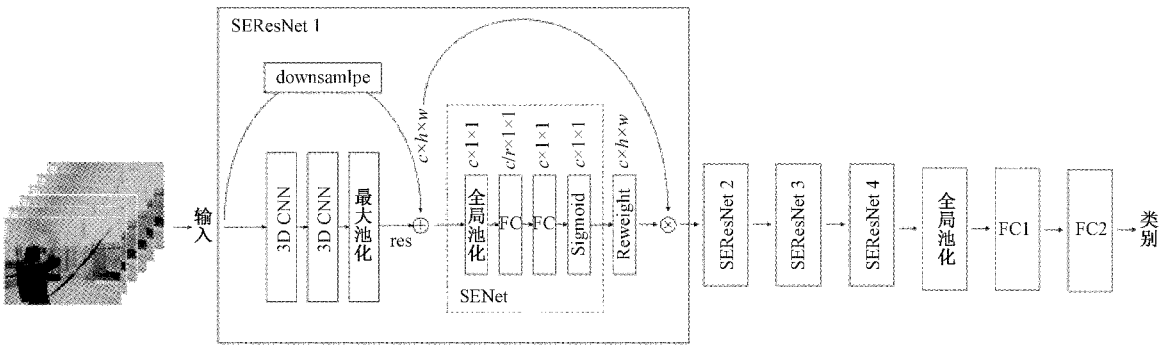

图 6　基于注意力机制的残差 3D 卷积网络

本文设计的网络模型具体参数如下。

SEResNet1 中，两个 3D 卷积核大小均为 3×3×3，个数均为 64；最大池化的内核大小为 1×2×2，步长为 1×2×2。SEResNet2 中 3D 卷积核个数为 128，其余参数与 SEResNet1 相同。

SEResNet3 与 SEResNet4 结构相同，两个 3D 卷积核大小均为 3×3×3，个数均为 256；最大池化的内核大小为 2×2×2，步长为 2×2×2。

全连接层 FC1 输入为 256，输出为 256，激活函数采用 ReLU 函数；全连接层 FC2 输入为 256，输出为类别数量，激活函数采用 ReLU 函数。

4　实验分析

4.1　数据集

本文的实验数据集采用公开的视频动作识别数据集 UCF-101 和 HMDB-51，部分动作的采样样本如图 7 所示。

图 7　数据集采样

HMDB-51 数据集包含有 51 个动作类别，共有 6 849 个视频片段。由于 HMDB-51 数据集里的视频片段大部分来自于电影片以及视频网站，所以其视频分辨率较低、视频帧所含噪声信息较多，导致在该数据集上的识别率较低，故而 HMDB-51 数据集更具挑战性。

UCF-101 数据集包含有 101 个动作类别，共有 13 320 个视频片段，其中每类动作由 25 个人做动作，每人做 4～7 组。

UCF-101 数据集是无约束的现实环境条件下拍摄的网络视频,在动作的采集上具有非常大的多样性,包括相机运行、外观变化、姿态变化、物体比例变化、背景变化、光纤变化等。UCF-101 数据库可以大致分为 5 类:人与物体互动、人体动作、人与人互动、乐器演奏、体育运动。

4.2　预处理与参数设置

在数据预处理阶段,首先将原始视频数据的视频帧大小调整为 128×171。对于视频分类来说,在 25～30 fps 的视频中,视频帧的采样间隔在 2～4 之间能够得到最好的效果,所以每间隔 4 帧保留一帧图像。最后选取连续且不重叠的 16 帧图像,对选取的视频帧图像进行抖动(随机剪裁)操作,最后的视频帧大小为 112×112。这种分辨率的输入兼顾准确率和计算效率,是在 GPU 显存受限下最优的输入分辨率。

预处理之后的视频帧数据进行归一化,最终得到的输入尺寸为(batch,3,16,112,112),其中 batch 为批处理大小,3 为图像通道数,16 为连续且不重叠的图像帧数。

本文的实验选择深度学习框架 Pytorch 作为实现平台,GPU 采用 GeForce RTX 2060 super。训练过程中,采用随机梯度下降方法(stochastic gradient descent, SGD),动量设置为 0.9。初始学习率设置为 0.001,每迭代 10 次后学习率缩小为原来的 1/10,直至迭代 100 次后停止训练。

4.3　实验结果分析

图 8 所示为在 UCF-101 数据集上在基于注意力机制的残差 3D 卷积网络模型的迭代过程。图中实线表示识别准确率,虚线表示损失值。可以看出在迭代次数达到 40 次时,动作识别的准确率已经超过 90%并且损失值迅速减小。随着迭代次数的增加,识别准确率和训练损失值的变化速率开始减小,并在接近迭代次数达到 60 次时开始趋于稳定,最终识别准确率的最高值为 95.8%。

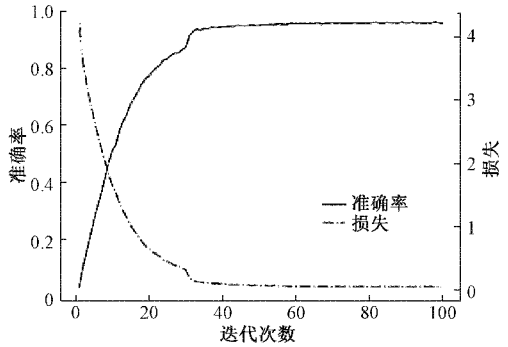

图 8　迭代过程

将本文方法与其他动作识别方法在 UCF-101 和 HMDB-51 数据集上进行比较,对比实验结果如表 1 所示。本文对比了改进的密集轨迹算法(iDT)、双流卷积神经网络(Two-Stream)、C3D 以及栅格化长短期记忆网络(Lattice LSTM)等主流方法。实验结果表明,本文提出的基于注意力机制和残差 3D 卷积网络的人体动作识别算法具有更好的识别效果。

表 1　实验结果对比　　%

算法	UCF-101	HMDB-51
iDT 算法	85.9	57.2
Two-Stream	88.0	59.4
C3D 算法	90.3	—
Lattice LSTM	93.6	66.2
本文算法	95.8	69.7

5　结　　论

本文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。首先利用残差 3D 卷积网络学习视频序列中连续视频帧之间的时空特征;之后利用扩展到三维结构的通道注意力网络对残差 3D 卷积结构学习到的每个特征通道赋予不同的权值;最后将重新标定权重的特征输入分类器得到最终的分类。在 UCF-101 和 HMDB-51 数据集上的实验表明,本文提出的动作识别方法可以很好地学习到连续视频帧之间的时间相关性,并且基于注意力机制能够更好的对时空特征加以利用。在接下来的工作中,要考虑如何在复杂度更高、时序相关性更强的视频动作中提高识别准确率。并且在保证动作识别准确率的情况下,尽量减少模型的参数量,从而提高模型的运行效率。

参考文献

[1]　钱慧芳,易剑平,付云虎. 基于深度学习的人体动作识别综述[J]. 计算机科学与探索, 2021, 15(3): 438-455.

[2]　YAN M J, MENG J J, ZHOU C L. Detecting spatiotemporal irregularities in videos via a 3D convolutional autoencoder [J]. Journal of Visual Communication and Image Representation, 2020, 15(12): 67-72.

[3]　KITANI K M, HUANG D A, MA W C. Activity forecasting [J]. Group & Crowd Behavior for Computer Vision, 2017, 17(10): 31-38.

[4]　ALEX K, ILYA S, GEOFFREY E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[5]　蓝金辉,王迪,申小盼. 卷积神经网络在视觉图像检测的研究进展[J]. 仪器仪表学报, 2020, 41(4): 167-182.

[6]　HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [J]. CVPR, 2016, 49(5): 770-778.

[7]　张培培,王昭,王菲. 基于深度学习的图像目标检测算

法研究[J]. 国外电子测量技术，2020，39(8)：34-39.
[8] LONG J，SHELHAMER E，DARRELL T. Fully convolutional networks for semantic segmentation[J]. CVPR，2015，42(10)：3431-3440.
[9] ZHANG Y，CUI X H，LIU Y，et al. Tire defects classification using convolution architecture for fast feature embedding [J]. International Journal of Computational Intelligence Systems，2018，11(1)：84-93.
[10] TRAN D，BOURDEV L，FERGUS R. Learning spatiotemporal features with 3D convolutional networks[C]. 2015 IEEE International Conference on Computer Vision(ICCV)，2015：4489-4497.
[11] DU T，RAY J，SHOU Z，et al. ConvNet architecture search for spatiotemporal feature learning[J]. Pattern Recognition，2017，45(9)：125-139.
[12] KOPERSKI M，BILINSKI P，BREMOND F. 3D trajectories for action recognition [C]. 2014 IEEE International Conference on Image Processing(ICIP)，2014：4176-4180.
[13] SIMONYAN K，ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Computer Vision and Image Understanding，2014，41(7)：568-576.
[14] 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.
[15] CHAUDHARI S，POLATKAN G，RAMANATH R. An attentive survey of attention models[J]. Image and Vision Computing，2019，75(2)：456-510.
[16] JADERBERG M，KAREN S，ANDREW Z. Spatial transformer networks[J]. Advances in Neural Information Processing Systems，2015，34(12)：2017-2025.
[17] HU J. Squeeze-and-excitation networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42(8)：2011-2023.
[18] ZHANG J，XIE Y，XIA Y，et al. Attention residual learning for skin lesion classification [J]. IEEE Transactions on Medical Imaging，2019，41(5)：510-527.

作者简介

龚捷，副院长，副教授，主要研究方向为计算机视觉、计算机网络。
E-mail：jieg@swpu.edu.cn

罗聪，工学硕士在读，主要研究方向为计算机视觉。
E-mail：luocongkyle@163.com

罗琴，工学博士，副教授，主要研究方向为计算机视觉。
E-mail：dorothy_lq@163.com