浙江理工大学学报,第 49 卷,第 3 期,2023 年 5 月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2023.03.004

基于全局时空注意力机制和 PCA_3DNet 的动作识别方法

田秋红,张元奎,潘　豪,李赛伟,施之翔
(浙江理工大学计算机科学与技术学院,杭州 310018)

摘　要:针对基于 3D 卷积神经网络的动作识别方法存在参数量过大、无法捕捉时空特征的全局依赖关系等问题,提出了一种基于全局时空注意力机制(Global spatiotemporal attention mechanism,GSTAM)和 PCA_3DNet 的动作识别方法。该方法引入伪 3D 卷积结构减少网络参数,在伪 3D 卷积结构中嵌入通道注意力机制(Channel attention mechanism,CAM)来增强通道特征,并采用全局时空注意力机制来捕捉特征信息的全局依赖关系,加强时空特征的表征能力,从而提高动作识别的准确率。该方法在两个公开数据集 UCF101 和 HMDB51 上的识别准确率分别为 93.5%和 70.5%,模型参数量为 13.46 Mi,浮点运算量为 8.73 Gi;在准确率、参数量和计算量上的综合表现优于现有的传统方法和深度学习方法。实验结果表明该方法能够获取丰富的时空特征信息,有效提升动作识别的性能。

关键词:全局时空注意力机制;PCA_3DNet;通道注意力机制;时空特征;动作识别
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1673-3851 (2023) 05-0310-08
引文格式:田秋红,张元奎,潘豪,等. 基于全局时空注意力机制和 PCA_3DNet 的动作识别方法[J]. 浙江理工大学学报(自然科学),2023,49(3):310-317.
Reference Format: TIAN Qiuhong, ZHANG Yuankui, PAN Hao, et al. Action recognition method based on global spatiotemporal attention mechanism and PCA_3DNet[J]. Journal of Zhejiang Sci-Tech University,2023,49(3):310-317.

Action recognition method based on global spatiotemporal attention mechanism and PCA_3DNet

TIAN Qiuhong, *ZHANG Yuankui*, *PAN Hao*, *LI Saiwei*, *SHI Zhixiang*
(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In view of the fact that the action recognition method based on three-dimensional (3D) convolutional neural network has the problems of too many parameters and cannot capture the global dependence of spatiotemporal features, an action recognition method based on global spatiotemporal attention mechanism (GSTAM) and PCA_3DNet is proposed. In this method, the pseudo 3D convolution structure is introduced to reduce network parameters, the channel attention mechanism (CAM) is embedded in the pseudo 3D convolution structure to enhance the channel features, and the GSTAM is adopted to capture the global dependence of feature information and strengthen the representation ability of spatiotemporal features, so as to improve the accuracy of action recognition. The recognition accuracy of this method on two public datasets UCF101 and HMDB51 is 93.5% and 70.5%, respectively, the amount of model parameters is 13.46 Mi, and the floating point of operations is 8.73 Gi. The comprehensive performance in accuracy, parameters and computation outperforms the existing traditional methods and

收稿日期:2022-10-10　　网络出版日期:2023-03-01
基金项目:国家自然科学基金项目(51405448);浙江省自然科学基金项目(LY20E050017)
作者简介:田秋红(1976—),女,辽宁兴城人,教授,博士,主要从事机器学习、模式识别、图像处理与识别方面的研究。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

deep learning methods. The experimental results show that the method can obtain abundant spatiotemporal feature information and effectively improve the performance of action recognition.

Key words: global spatiotemporal attention mechanism; PCA_3DNet; channel attention mechanism; spatiotemporal feature; action recognition

0 引　言

动作识别在智能视频监控[1]、运动分析、智能人机交互等领域有着广泛的应用前景[2],已经逐渐成为一个非常热门且具有挑战性的研究方向。目前动作识别方法主要分为传统动作识别方法和基于深度学习的动作识别方法[3]。传统动作识别方法主要通过手工提取视频动作的运动特征。Wang 等[4]提出密集轨迹(Dense trajectories, DT)算法来获取视频动作的运动轨迹,提取方向梯度直方图(Histogram of oriented gradient, HOG)[5]、光流方向直方图(Histograms of oriented optical flow, HOF)[6]特征。许培振等[7]对 DT 算法进行改进,提出了改进的密集轨迹(Improved dense trajectories, IDT)算法,该算法通过加速稳健特征(Speeded-up robust features, SURF)匹配算法来获取视频帧之间的光流特征。Patel 等[8]利用运动目标检测和分割,提取出运动对象的 HOG 特征,并融合速度、位移及区域特征来表征动作。Xia 等[9]对 IDT 的光流轨迹算法进行了扩展,设计了一种多特征融合的描述子表示动作。传统动作识别方法的局限在于动作识别的准确率较低,手工提取特征不够充分,并且计算成本较大。

随着深度学习技术的迅速发展,越来越多的研究人员利用卷积神经网络[10]自动提取图像特征。Simonyan 等[11]提出了一种双流动作识别网络,该网络通过空间流网络和时间流网络来提取外观特征和运动特征,但是该网络主要考虑外观和短期运动,不利于建模时间跨度较大的视频任务。Wang 等[12]提出了一种时间分段网络(Temporal segment networks, TSN)来弥补双流网络中建模长时间视频动作的不足;Wang 等[13]又对 TSN 网络进行了改进,提出了一种能够捕获多尺度时间信息的时间差异网络(Temporal difference networks, TDN)。虽然上述方法[11-13]能够提取出视频中动作的时间特征和空间特征,但是这些方法在时空特征提取上是相互独立的。Tran 等[14]使用三维卷积网络(Convolutional 3d networks, C3D)来直接学习视频中动作的时空特征。Carreira 等[15]将 InceptionV1 网络中所有二维(Two-dimensional, 2D)卷积全部膨胀成三维(Three-dimensional, 3D)卷积,提出了膨胀三维卷积网络(Inflated 3d convolution networks, I3D)。Hara 等[16]将 3D 卷积应用到残差网络上,提出了三维残差网络。Qiu 等[17]提出了一种伪三维卷积网络(Pseudo-3D convolution networks, P3D),该网络通过伪 3D 卷积结构来拟合 3D 卷积,从而缓解了 3D 卷积导致模型参数量过大的问题,并且实验验证了伪 3D 卷积结构的有效性。上述研究人员采用了多种 3D 卷积神经网络用于动作识别,但是在使用 3D 卷积进行特征提取的过程中,无法区分关键动作特征和背景特征,且在卷积过程当中无法获取特征的全局依赖关系。

注意力机制被引入卷积神经网络之后能够显著提升网络的性能,使得网络关注图像中关键的信息,抑制无关信息[18]。Wang 等[19]将一种残差注意网络应用于图像分类任务上,并取得较好的分类效果。Jaderberg 等[20]提出了一种空间注意力机制,将原始图像的空间信息转换到另一个空间,保留其关键信息,结果表明该方法能够有效提高模型性能。Hu 等[21]提出了一种通道注意力模型 SeNet,通过对输入特征图的通道赋予不同的注意力权重来学习不同通道特征的重要性。Woo 等[22]结合上述两个方法提出了一种卷积块注意力模型(Convolution block attention module, CBAM),该模型由通道注意力模块和空间注意力模块构成。Lei 等[23]提出了一种通道式时间注意力网络,该网络通过通道注意力来强调每一帧的细粒度信息特征,且实验证明了该注意力可以提升网络模型的表达能力。虽然上述研究方法在动作识别的任务中取得了一定的效果,但是仍然没有考虑到时空特征的全局依赖关系。

本文针对动作识别方法中存在的特征提取不充分、参数量过多、无法捕获时空特征的全局依赖关系等问题,提出了一种基于全局时空注意力机制和 PCA_3DNet 的动作识别方法。为了减少模型的参数量,本文引入伪 3D 卷积结构代替 3D 卷积结构,采用串联 1×1×3 卷积和 3×3×1 卷积的方式来拟合 3×3×3 卷积,以减少参数量;为了充分利用动作特征的通道信息,将通道注意力机制嵌入伪 3D 卷积结构中,实现通道特征信息的增强,并设计了

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

PCA_3DNet网络模型作为特征提取网络;加强时空特征的表征能力,将全局时空注意力机制加入PCA_3DNet网络中,对时空特征的全局依赖关系进行建模,以提高视频动作特征的提取能力。

1 方法设计

本文设计了一种基于全局时空注意力机制和PCA_3DNet的动作识别网络模型,该模型整体框架示意图如图1所示。首先采用稀疏采样的方法提取视频帧序列作为模型的输入;然后通过PCA_3D卷积block提取视频动作特征,并加入全局时空注意力机制学习时空特征的全局依赖关系,使模型提取到更丰富的动作特征;最后使用Softmax层实现动作识别。

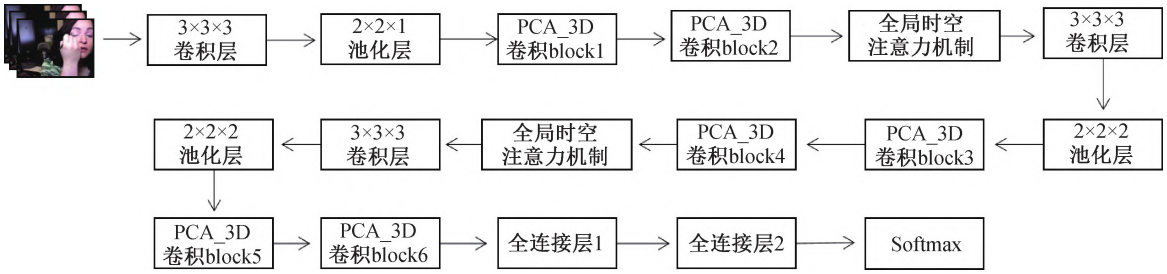

图1 网络模型整体框架示意图

1.1 PCA_3DNet

本文通过PCA_3D卷积block构建了特征提取网络——PCA_3DNet,其包含6个PCA_3D卷积block,PCA_3D卷积block结构示意图如图2所示。

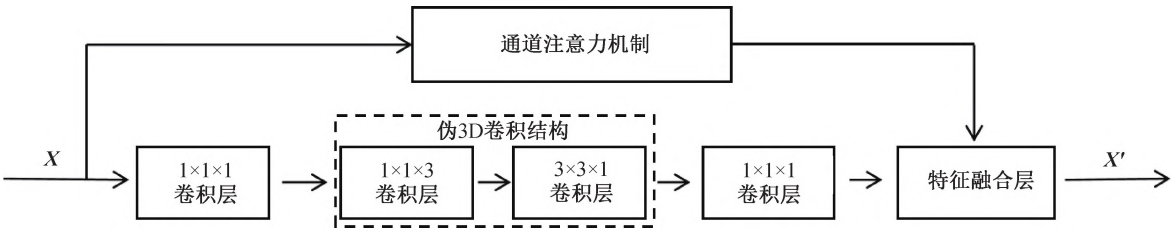

图2 PCA_3D卷积block结构示意图

首先基于3D卷积层对输入特征$\boldsymbol{X}\in\mathbf{R}^{H\times W\times T\times C}$提取时空特征,其中:$H$、$W$、$T$和$C$分别表示特征图的高度、宽度、时间深度和通道数。在PCA_3D卷积block中,通过伪3D卷积结构[17](伪3D卷积结构由1×1×3卷积层和3×3×1卷积层构成)来模拟3×3×3卷积提取时空特征,以减少参数量。3D卷积层参数量的计算公式为:$(k_{\mathrm{h}}\times k_{\mathrm{w}}\times k_{\mathrm{t}}\times n_{\mathrm{ic}}+1)\times n_{\mathrm{oc}}$,其中:$k_{\mathrm{h}}$、$k_{\mathrm{w}}$、$k_{\mathrm{t}}$为3D卷积核在高、宽、时间三个维度的大小,$n_{\mathrm{ic}}$为输入特征图的通道数量,$n_{\mathrm{oc}}$为3D卷积核的数量。其次,本文在PCA_3D卷积blcok中嵌入通道注意力机制(Channel attention mechanism,CAM)模块,该模块针对输入特征$\boldsymbol{X}$的通道关系进行建模,能够获取特征的通道信息权重分布,加强有用通道特征,抑制无关通道特征,从而增强PCA_3D卷积block的特征提取能力。最后利用特征融合层将1×1×1卷积层的输出特征和CAM模块的输出特征相融合,得到PCA_3D卷积block的输出特征$\boldsymbol{X}'$。其中在PCA_3D卷积block中嵌入的通道注意力机制结构示意图如图3所示。

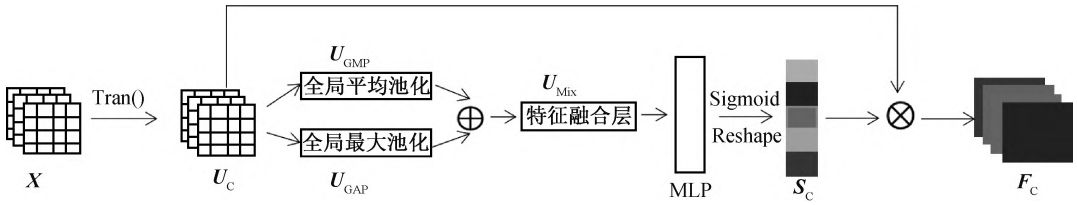

图3 通道注意力机制结构示意图

CAM模块首先利用特征线性转换层将输入特征$\boldsymbol{X}\in\mathbf{R}^{H\times W\times T\times C}$转换成$\boldsymbol{U}_{\mathrm{C}}\in\mathbf{R}^{H\times W\times T\times C}$;其次基于全局平均池化操作和全局最大池化操作将$\boldsymbol{U}_{\mathrm{C}}\in\mathbf{R}^{H\times W\times T\times C}$压缩为$\boldsymbol{U}_{\mathrm{GAP}}\in\mathbf{R}^{1\times1\times1\times C}$和$\boldsymbol{U}_{\mathrm{GMP}}\in\mathbf{R}^{1\times1\times1\times C}$,生成两个不同的通道特征描述符$\boldsymbol{U}_{\mathrm{GAP}}$和$\boldsymbol{U}_{\mathrm{GMP}}$。通道特征描述符$\boldsymbol{U}_{\mathrm{GAP}}$和$\boldsymbol{U}_{\mathrm{GMP}}$对输入特征的全局像素进行计算,因此$\boldsymbol{U}_{\mathrm{GAP}}$和$\boldsymbol{U}_{\mathrm{GMP}}$具有全局时空特征的感受野。上述计算过程可用式(1)—(3)表示:

$$\boldsymbol{U}_{\mathrm{C}}=\mathrm{Tran}(\boldsymbol{X}) \tag{1}$$

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\boldsymbol{U}_{\mathrm{GAP}}=\mathrm{GAP}(\boldsymbol{U}_{\mathrm{C}})=\frac{1}{H\times W\times T}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{T}\boldsymbol{U}_{\mathrm{C}}(i,j,k)\quad(2)$$

$$\boldsymbol{U}_{\mathrm{GMP}}=\mathrm{GMP}(\boldsymbol{U}_{\mathrm{C}})=\max\{\boldsymbol{U}_{\mathrm{C}}(i,j,k)\},\ i=1,\cdots,H;j=1,\cdots,W;k=1,\cdots,T\quad(3)$$

其中:Tran( )表示在输入特征 $\boldsymbol{X}$ 上的相应通道上进行 1×1×1 卷积线性转换;GAP()表示全局平均池化操作,GMP()表示全局最大池化操作;$\boldsymbol{U}_{\mathrm{C}}$ 表示经过 1×1×1 卷积层线性转换后的输出特征;i、j、k 分别表示在特征图 $\boldsymbol{U}_{\mathrm{C}}$ 中 H、W、T 维度上的位置。

为了学习更为丰富的通道特征,本文选择使用特征融合层来融合两个通道特征描述符 $\boldsymbol{U}_{\mathrm{GAP}}$ 和 $\boldsymbol{U}_{\mathrm{GMP}}$,生成混合通道特征描述符 $\boldsymbol{U}_{\mathrm{Mix}}\in\mathbf{R}^{1\times1\times1\times C}$;随后将混合通道特征描述符 $\boldsymbol{U}_{\mathrm{Mix}}$ 输入到 MLP(多层感知机),并经过 Sigmoid 和 Reshape 操作生成通道特征相关性描述符 $\boldsymbol{S}_{\mathrm{C}}\in\mathbf{R}^{1\times1\times1\times C}$;最后,将通道特征相关性描述符 $\boldsymbol{S}_{\mathrm{C}}$ 和特征 $\boldsymbol{U}_{\mathrm{C}}$ 逐通道相乘,得到通道注意力特征 $\boldsymbol{F}_{\mathrm{C}}\in\mathbf{R}^{H\times W\times T\times C}$。上述计算过程可用式(4)—(6)表示:

$$\boldsymbol{U}_{\mathrm{Mix}}=\boldsymbol{U}_{\mathrm{GMP}}\oplus\boldsymbol{U}_{\mathrm{GAP}}\quad(4)$$

$$\boldsymbol{S}_{\mathrm{C}}=\sigma(\mathrm{MLP}(\boldsymbol{U}_{\mathrm{Mix}}))\quad(5)$$

$$\boldsymbol{F}_{\mathrm{C}}=\boldsymbol{S}_{\mathrm{C}}\otimes\boldsymbol{U}_{\mathrm{C}}\quad(6)$$

其中:$\boldsymbol{F}_{\mathrm{C}}$ 是 CAM 模块的输出特征,⊗表示矩阵相乘,σ 表示 sigmoid 函数操作,⊕表示特征融合操作。

1.2　全局时空注意力机制模块

在 3D 卷积过程中,输出特征的目标特征,是由输入特征和卷积核在感受野范围内进行局部内积运算得到,所以 3D 卷积在特征提取的过程中仅仅考虑到了输入特征的局部信息。3D 卷积过程示意图如图 4 所示。

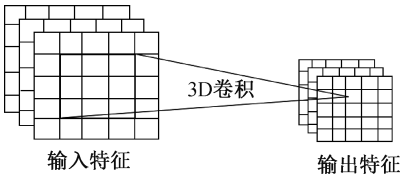

图 4　3D 卷积过程示意图

图 4 中输出特征的目标特征仅仅作用于输入特征的局部区域,然而对于输入数据是视频帧序列,目标特征不仅仅依赖于输入特征中的局部特征信息,还可能依赖于其他时空特征信息。因此本文提出了全局时空注意力机制(Global spatiotemporal attention mechanism, GSTAM)模块,该模块通过计算当前特征位置和其他时空特征位置的相关性来捕获时空特征之间的全局依赖关系,全局时空注意力机制结构示意图如图 5 所示。

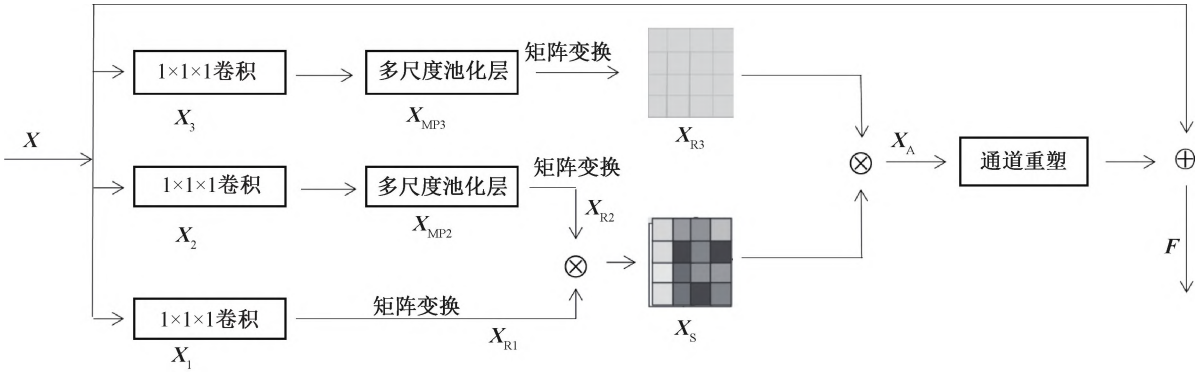

图 5　全局时空注意力机制结构示意图

GSTAM 模块首先将输入特征 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times T\times C}$ 分别经过 3 个 1×1×1 卷积,得到 $\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3\in\mathbf{R}^{H\times W\times T\times C}$。其次,将特征图 $\boldsymbol{X}_2$ 和特征图 $\boldsymbol{X}_3$ 输入多尺度池化层,多尺度池化层结构示意图如图 6 所示,通过多尺度池化层对 $\boldsymbol{X}_2$、$\boldsymbol{X}_3$ 进行降采样操作,得到 $\boldsymbol{X}_{\mathrm{MP2}}$、$\boldsymbol{X}_{\mathrm{MP3}}\in\mathbf{R}^{H_1\times W_1\times T_1\times C}$,其中:$H_1$、$W_1$、$T_1$ 分别表示特征图的高度、宽度和时间深度。然后对特征图 $\boldsymbol{X}_1$、$\boldsymbol{X}_{\mathrm{MP2}}$、$\boldsymbol{X}_{\mathrm{MP3}}$ 进行矩阵变换得到 $\boldsymbol{X}_{\mathrm{R1}}\in\mathbf{R}^{N\times C}$($N=H\times W\times T$)、$\boldsymbol{X}_{\mathrm{R2}}\in\mathbf{R}^{C\times S}$($S=H_1\times W_1\times T_1$)、$\boldsymbol{X}_{\mathrm{R3}}\in\mathbf{R}^{S\times C}$($S=H_1\times W_1\times T_1$);将 $\boldsymbol{X}_{\mathrm{R1}}$ 和 $\boldsymbol{X}_{\mathrm{R2}}$ 进行矩阵相乘计算当前特征位置和其他特征之间的相关性,并通过 Softmax 函数生成全局时空注意力权重系数 $\boldsymbol{X}_{\mathrm{S}}\in\mathbf{R}^{N\times S}$。在得到全局时空注意力权重系数后,将其和 $\boldsymbol{X}_{\mathrm{R3}}$ 进行逐元素相乘得到包含注意力的特征图 $\boldsymbol{X}_{\mathrm{A}}\in\mathbf{R}^{N\times C}$,之后将 $\boldsymbol{X}_{\mathrm{A}}$ 进行通道重塑后和输入特征 $\boldsymbol{X}$ 进行残差连接,得到 GSTAM 模块的输出特征 $\boldsymbol{F}\in\mathbf{R}^{H\times W\times T\times C}$。

多尺度池化层结构由池化核大小分别为 2、4、8 的最大池化层组成,通过多尺度池化层结构能够从多维度压缩特征,提取出不同尺度的池化特征,使得网络能够学习到不同尺度下的特征信息,并且多尺度池化层结构降低了 GSTAM 模块中特征图的大小,从而减少了矩阵相乘产生的较大计算量。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

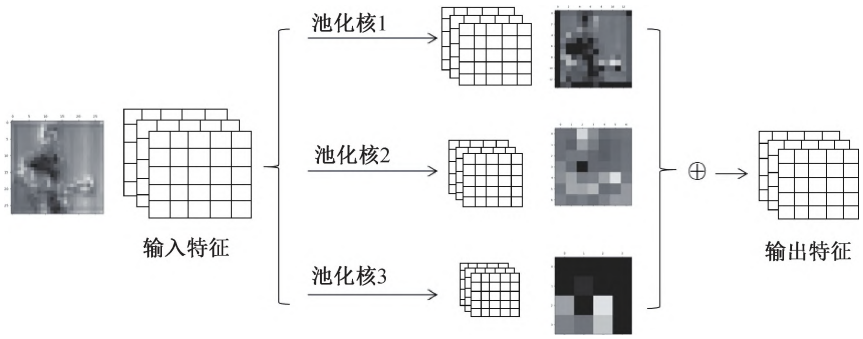

图6 多尺度池化层结构示意图

2 实验与结果分析

本文在UCF101和HMDB51这两个具有挑战性的动作识别数据集上测试本文提出的方法,并且从不同的角度来验证本文提出方法的有效性和可行性。

2.1 数据集

UCF101数据集:该数据集是一个真实动作视频集,该数据集包含101类动作,一共有13320个视频片段,每个类别的视频动作分为25组,每组包含4～7个视频动作,视频类别主要分类5类,分别是人与物体交互、人体动作、人与人交互、乐器演奏、体育运动[24]。部分示例视频截图如图7所示。

HMDB51数据集:该数据集包含了51类动作,共有6849个视频片段。视频类别主要分为面部动作、面部操作、身体动作、交互动作、人体动作等5类,如抽烟、拍手、打球、拥抱等动作,该数据集的视频大多来源于电影剪辑片段,小部分来源于YouTube等视频网站,像素较低[25]。部分示例视频截图如图8所示。

图7 UCF101数据集示例视频截图

图8 HMDB51数据集示例视频截图

2.2 实验过程

本文实验基于Python3.7、Tensorflow2.0、Keras2.0实现,选择稀疏采样的方法从视频片段中提取视频帧作为模型的输入,在UCF101数据集上分别选取8、12、16帧视频帧作为模型输入进行了实验,实验结果如表1所示。根据实验确定本文网络模型输入大小设置为112×112×16×3,采用Adam()优化器学习网络参数,batch大小设置为16,初始学习率设置为0.001,权重衰减设置为0.005,防止过拟合添加的Dropout层的失活率设置为0.5,模型训练迭代次数达到150次后终止训练。

2.3 消融实验

为了验证在PCA_3D卷积block中嵌入的CAM模块和在PCA_3DNet中添加的GSTAM模块的可行性和有效性,本文在UCF101数据集上进行了消融实验。

表1 UCF101数据集不同视频帧数的实验结果

帧数/帧	准确率/%
8	88.93
12	90.36
16	93.50

本文分别使用P3D卷积结构搭建的Baseline模型、Baseline+CAM(PCA_3DNet)模型、Baseline+GSTAM模型、Baseline+CAM+GSTAM(本文方法)在UCF101数据集上进行实验。实验结果见图9。从图9(a)中可以看出,当本文分别在

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Baseline 模型的基础上添加 CAM 模块、GSTAM 模块后,模型的准确率都比 Baseline 模型高,说明 CAM 模块和 GSTAM 模块可以增强模型的特征提取能力,提升模型性能。当本文将 CAM 模块和 GSTAM 模块都添加到 Baseline 模型当中,本文方法的准确率比 Baseline 模型、Baseline+CAM 模型和 Baseline+GSTAM 模型都高,说明加入 CAM 模块和 GSTAM 模块后,本文方法可以提取到更加丰富的特征,使得网络模型的性能显著提升。从图 9(b)中可以看出,在前 30 次迭代的时候,各组实验模型的损失值都下降最快;在迭代到 100 次以后,各组实验模型的损失值都趋于稳定,而本文所采取的 Baseline+CAM+GSTAM 模型的波动性最小,损失值更加稳定。实验结果表明:CAM 模块能够有效增强通道特征,GSTAM 模块能够学习时空特征的全局依赖关系,添加两个模块能够增强模型的特征提取能力,有效提升网络模型的识别准确率。

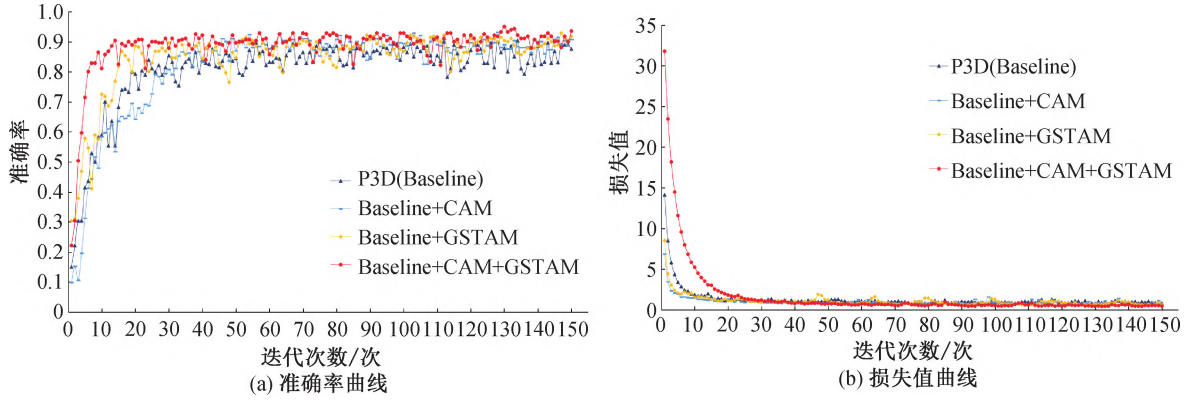

图 9　UCF101 数据集上不同模块的实验准确率曲线和损失值曲线

本文同时在参数量和浮点运算量(Floating point operations, FLOPs)方面来评估 CAM 模块和 GSTAM 模块的有效性,其中 FLOPs 可以表示为计算量,用于衡量模型方法的复杂度。实验结果见表 2。从表 2 中可以看出,Baseline 模型的浮点运算量为 8.53 Gi,参数量为 13.19 Mi。在分别添加了 CAM 模块和 GSTAM 模块后,模型的浮点运算量的增量以及参数量的增量非常少,但是准确率上的提升较为明显,意味着本文以较小的内存代价、计算量代价换取了准确率较大的提升,并且本文方法的浮点运算量的增量为 0.20 Gi,参数量的增量为 0.27 Mi,准确率却提升了 5.96%。实验结果表明:本文提出的 CAM 模块以及 GSTAM 模块可行并且有效,能够提升模型的性能且花费的计算成本较低。

表 2　添加不同模块的准确率、参数量、浮点运算量实验结果对比

方法	是否添加 CAM 模块	是否添加 GSTAM 模块	准确率/%	浮点运算量/Gi	浮点运算量增量/Gi	参数量/Mi
Baseline	否	否	87.54	8.53	—	13.19
Baseline+CAM	是	否	90.75	8.54	0.01	13.42
Baseline+GSTAM	否	是	91.77	8.58	0.05	13.42
本文方法	是	是	93.50	8.73	0.20	13.46

2.4　方法对比

为了验证本文方法的可行性,本文将本文方法与主流方法在 UCF101 数据集和 HMDB51 数据集上进行对比实验,实验结果见表 3 所示。从表 3 中可以看出,在 UCF101 数据集上,除了 I3D(Two-Stream)[15]和 TSN(RGB+Flow)[12]外,本文方法和其他方法相比都显示出了优势。而 I3D(Two-Stream)模型和 TSN(RGB+Flow)模型都需要进行光流的计算,这会增加模型的计算的复杂度,并且影响模型的实时性能。在 HMDB51 数据集上,本文方法和其他方法相比,本文方法准确率最高,性能表现最佳。实验结果表明:本文方法在不同的数据集上都具备较高的准确率,模型泛化能力强,鲁棒性高。

此外,本文方法的准确率比 C3D 模型高了 7.7%;与 P3D 模型相比提高了 4.9%;与 Two-Stream 模型相比提高了 5.5%。虽然 I3D(Two-Stream)模型和 TSN(RGB+Flow)模型在 UCF101 数据集上的准确率比本文方法高 0.2%和 0.7%,但是这两种方法引入了双流结构并将光流数据作为输入,而本文方法仅需 RGB 数据作为输入,减少了光流数据的计算成本。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表3 不同方法在UCF101数据集和HMDB51数据集上的准确率

方法	预训练数据集	准确率/%	
		UCF101	HMDB51
HOG[5]	—	72.4	40.2
C3D[14]	Sport-1M	85.8	54.9
P3D[17]	Kinetics	88.6	—
T3D[26]	—	93.2	—
Two-Stream[11]	—	88.0	—
Method[23]	ImageNet+Kinetics	91.1	—
ARTNet[27]	Kinetics	93.5	67.6
ResNeXt-101[28]	—	90.7	61.0
Method[29]	—	91.5	67.9
I3D(Two-Stream)[15]	—	93.7	—
TSN(RGB)[12]	—	85.1	51.0
TSN(RGB+Flow)[12]	—	94.2	69.4
本文方法	—	93.5	70.5

为了进一步验证本文方法的有效性,本文和主流方法在浮点运算量和参数量上进行了对比实验,实验结果见表4。从表3—表4中可以看出,虽然I3D(Two-Stream)在准确率上面比本文方法高0.2%,但是该模型是直接将Inception V1中的2D卷积膨胀成3D卷积,从而在参数量方面远远超过本文方法,说明该模型需要耗费更多的内存代价;TSN(RGB+Flow)模型虽然在UCF10数据集准确率比本文方法高0.7%,但是该方法的浮点运算量为16 Gi,约是本文方法的两倍,并且TSN模型在使用RGB数据作为输入的时候,准确率比本文方法低8.4%。本文方法与C3D模型和P3D模型相比,参数量约为C3D模型的1/6、P3D模型的1/5。在浮点运算量方面,本文方法的浮点运算量较小,说明本文方法的模型复杂度低,和其他方法相比,本文方法也具备优势。实验结果表明本文方法在模型准确率、模型参数量、模型计算量等方面取得了较好的平衡,在具有较低的参数量和计算量的同时能够拥有较高的准确率。

表4 不同方法的浮点运算量和参数量实验结果对比

方法	浮点运算量/Gi	参数量/Mi
ResNeXt-101[28]	9.68	—
I3D(Two-Stream)[15]	—	250.00
TSN(RGB+Flow)[12]	16.00	—
T3D[26]	19.80	85.50
C3D[14]	71.82	78.40
P3D[17]	—	63.70
本文方法	8.73	13.46

3 结 论

本文提出了一种基于全局时空注意力机制和PCA_3DNet的动作识别方法。该方法通过搭建PCA_3DNet作为主干特征提取网络,并且利用其内部的PCA_3D卷积block结构来减少网络参数以及增强通道特征信息;通过全局时空注意力机制模块可以获取特征的全局依赖关系,进一步提升特征提取效率,从而提高动作识别的准确率。本文方法在UCF101和HMDB51公开数据集上进行了实验,识别准确率分别为93.5%和70.5%,参数量为13.46 Mi,浮点运算量为8.73 Gi;消融实验证明了本文方法能够提取到更加丰富的时空特征,在动作识别任务中可以实现更好的性能;对比实验证明了本文方法的准确率较高,参数量和计算量较少且具有较高的鲁棒性。目前本文在公开的动作数据集上进行实验,后续将采集实际场景下的动作视频数据集,并对现有方法的网络结构进行优化,以适用于实时场景下的动作识别任务。

参考文献:

[1] Ben Mabrouk A, Zagrouba E. Abnormal behavior recognition for intelligent video surveillance systems[J]. Expert Systems with Applications: An International Journal, 2018, 91(C):480-491.

[2] Wang L, Huynh D Q, Koniusz P. A comparative review of recent kinect-based action recognition algorithms[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2020, 29: 15-28.

[3] 卢修生,姚鸿勋. 视频中动作识别任务综述[J]. 智能计算机与应用, 2020, 10(3): 406-411.

[4] Wang H, Kläser A, Schmid C, et al. Action recognition by dense trajectories[C]//CVPR. Colorado Springs, CO, USA. IEEE, 2011: 3169-3176.

[5] Klaeser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients[C]//BMVC 2008-19th British Machine Vision Conference. Leeds. British Machine Vision Association, 2008: 1-10.

[6] Brox T, Malik J. Large displacement optical flow: descriptor matching in variational motion estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 500-513.

[7] 许培振, 余志斌, 金炜东, 等. 基于提高的稠密轨迹人体行为识别[J]. 系统仿真学报, 2017, 29(9): 2053-

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2058.

[8] Patel C I, Labana D, Pandya S, et al. Histogram of oriented gradient-based fusion of features for human action recognition in action video sequences[J]. Sensors, 2020, 20(24):7299.

[9] Xia L M, Ma W T. Human action recognition using high-order feature of optical flows[J]. The Journal of Supercomputing, 2021, 77(12): 14230-14251.

[10] Heslinga F G, Pluim J P W, Dashtbozorg B, et al. Approximation of a pipeline of unsupervised retina image analysis methods with a CNN[C]// Medical Imaging 2019: Image Processing. San Diego, USA. SPIE, 2019, 10949: 416-422.

[11] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. New York. ACM, 2014: 568-576.

[12] Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks for action recognition in videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740-2755.

[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, TN, USA. IEEE, 2021: 1895-1904.

[14] 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. IEEE, 2016: 4489-4497.

[15] 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, HI, USA: IEEE, 2017: 4724-4733.

[16] Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3d residual networks for action recognition [EB/OL]. (2017-08-25)[2022-10-10]. https://arxiv.org/abs/1708.07632.

[17] Qiu Z F, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy. IEEE, 2017: 5534-5542.

[18] 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.

[19] Wang F, Jiang M Q, Qian C, et al. Residual attention network for image classification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 6450-6458.

[20] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[EB/OL]. (2015-06-05)[2022-10-10]. https://arxiv.org/abs/1506.02025.

[21] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 7132-7141.

[22] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]// Proceedings of the European conference on computer vision (ECCV). Cham: Springer International Publishing, 2018: 3-19.

[23] Lei J J, Jia Y L, Peng B, et al. Channel-wise temporal attention network for video action recognition[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China. IEEE, 2019: 562-567.

[24] Soomro K, Zamir A R, Shah M. A dataset of 101 human action classes from videos in the wild[EB/OL]. (2012-12-03)[2022-10-10]. https://arxiv.org/abs/1212.0402.

[25] Wishart D S, Tzur D, Knox C, et al. HMDB: the human metabolome database[J]. Nucleic Acids Research, 2007, 35(suppl_1): D521-D526.

[26] Liu K, Liu W, Gan C, et al. T-C3D: Temporal convolutional 3D network for real-time action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1):7138-7145.

[27] Wang L M, Li W, Li W, et al. Appearance-and-relation networks for video classification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 1430-1439.

[28] 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, UT, USA. IEEE, 2018: 6546-6555.

[29] Xu J, Song R, Wei H L, et al. A fast human action recognition network based on spatio-temporal features[J]. Neurocomputing, 2021, 441: 350-358.

(责任编辑:康　锋)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net