高技术通讯 2024 年 第 34 卷 第 6 期:590-601

doi:10.3772/j.issn.1002-0470.2024.06.004

时空多尺度关联特征融合的二维卷积网络细粒度动作识别模型①

胡正平②[*,**] 王昕宇[*] 董佳伟[*] 赵艳霜[*] 刘 洋[*]

([*]燕山大学信息科学与工程学院 秦皇岛 066004)

([**]燕山大学河北省信息传输与信号处理重点实验室 秦皇岛 066004)

摘 要 针对传统二维(2D)卷积网络提取时空特征尺度单一以及对细粒度动作数据集中帧与帧之间的远程时间关联信息利用不足的问题,本文提出时空多尺度关联特征融合的 2D 卷积网络细粒度动作识别模型。首先,为建模视频多尺度空间关联以加强对细粒度视频数据的空间表征能力,模型使用多尺度"特征压缩、特征激发"方式,使网络所提取空间特征更加丰富有效。然后,为充分利用细粒度视频数据时间维度上的运动信息,本文引入时间窗口自注意力机制,利用自注意力机制强大的远程依赖建模能力同时只在时间维度上进行自注意力操作,以较低计算成本建模远程时间依赖关系。最后,考虑到所提取时空特征对不同类型动作分类的贡献不均等,本文引入自适应特征融合模块,为特征动态赋予不同权重实现自适应特征融合。模型在 2 个细粒度动作识别数据集 Diving48 和 Something-somethingV1 上识别准确率分别达到 86.0% 和 46.9%,分别使原始主干网络识别准确率提升 3.8% 和 1.3%。实验结果表明,在只使用视频帧信息作为输入的情况下,本模型达到与现有基于 Transformer 和三维卷积神经网络(3D CNN)算法相当的识别准确率。

关键词 细粒度动作识别;多尺度时空关联特征;远程依赖建模;自注意力机制

基于视频的动作识别由于在视频监控、医疗诊断、军事安防等方面有广泛应用前景,已经成为视频理解领域研究热点。相比于图像识别,视频动作识别难度更大,其中关键因素在于引入时间维度信息,尤其是当涉及细粒度动作识别时,问题则变得更具有挑战性。细粒度动作识别要求模型能够捕获相似动作在时间和空间维度上微小差异,但是经典模型却只专注于视频数据中空间信息,对于如何有效提取视频数据中时间和空间特征,目前研究思路主要包括:(1)提前抽取光流图像用来建模运动信息;(2)构建基于三维(3-dimension,3D)卷积的模型直接提取时空特征;(3)使用 Transformer 和自注意力机制提取远距离时空关联特征。

早期动作识别方法直接使用二维卷积网络对单个视频帧中静态信息建模,缺点是对视频信息利用效率不足。直到 2014 年,Simonyan 和 Zisserman[1] 设计了基于二维卷积神经网络(2-dimensional convolutional neural network,2D CNN)的双流网络。双流网络有 2 个分支:其中一个分支输入视频帧,用于建模空间信息;另一个分支输入光流图像,用于建模视频中运动信息。由于双流网络成功捕获运动信息,因此识别精度相比于早期方法有较大提升。为解决双流网络对关联信息不敏感问题,Yue-Hei 等人[2] 在双流网络 2 个分支中加入长短期记忆网络(long short term memory,LSTM)来捕获关联信息;由于双流网络只在网络最后对 2 个分支分类结果融合,对所提取时空信息利用不够充分,于是 Feichtenhofer 等人[3] 对双流网络中 2 个分支如何进行融合进行有

① 国家自然科学基金面上项目(61771420)和国家自然科学基金(62001413)资助项目。

② 男,1970 年生,博士,教授;研究方向:深度视频学习模型;联系人,E-mail: hzp@ysu.edu.cn。

(收稿日期:2023-11-08)

益的探索,发现在网络早期阶段对 2 个分支浅层特征进行融合可以提高双流网络识别精度。为提高双流网络处理长视频的能力,2016 年 Wang 等人[4]在双流网络基础上构建时间分段网络(temporal segment networks,TSN)模型,TSN 将一个长视频切分成几段,分别将每段视频输入到一组参数共享的双流网络中进行分类,取得了较好效果。双流网络使用光流图像来提取视频中运动信息,相比于以往方法识别精度较高,但它需要提前抽取光流图像,这个步骤需要耗费较长时间。不同于 2D 卷积,3D 卷积可以同时提取空间和时间信息,更适合处理视频信息。在 2015 年,Tran 等人[5]使用 3D 卷积网络(convolutional 3D,C3D)进行动作识别,从此 3D 卷积网络开始成为视频理解领域主流模型。为解决 C3D 网络参数众多难以训练的问题,Carreira 等人[6]在 2017 年提出双流膨胀 3D 卷积网络网络(inflated 3D ConvNet,I3D),将在 ImageNet 上预训练的 2D 卷积网络模型扩充为 3D 模型,充分利用 2D 模型的预训练参数。为进一步提高 3D 卷积网络识别精度,同时受到双流网络启发,在 2018 年,Feichtenhofer 等人[7]参考双流网络结构,构建了 SlowFast 网络。SlowFast 使用快速 3D 网络分支学习运动信息、慢速 3D 网络分支学习静态信息,取得较高识别准确率。虽然 3D 卷积能同时捕获时空信息,但是受限于卷积核尺寸,不利于提取远距离关联信息。自注意力机制和 Transformer 有较强远程关联建模能力,一经提出就在各个领域得到关注。2018 年,Wang 等人[8]提出非局域(Non-Local)算子,此方法类似于自注意力机制,用于长距离关联建模,开创式地将自注意力机制引入到视频理解领域。不同于 Non-Local 算子,Zhang 等人[9]在 2021 年提出时间查询网络(temporal query networks,TQN)将 3D 卷积网络主干和 Transformer 解码器结合起来实现细粒度动作识别。TQN 使用可分离 3D 卷积(separable 3D convolutions,S3D)网络作为视觉主干提取特征,特征与查询向量输入到 Transformer 解码器中,最终输出相应响应向量。考虑到这些研究存在特征尺度单一问题,Bera[10]在 2023 年提出基于图像块的注意力机制(patch-based attention,PbA),利用多尺度注意力促使 CNN 模型理解上下文语义信息。为在视频理解领域完全脱离卷积操作,Bertasius 等人[11]在 2021 年提出实现完全使用自注意力机制的动作识别网络,作者对比多种形式注意力计算方法,得出最优计算形式。由于视频领域 Transformer 网络参数过多,为降低计算负担,Liu 等人[12]在 2022 年提出视频分层视觉 Transformer(video swin Transformer),将图像领域 swin Transformer 中窗口自注意力机制拓展到视频领域,降低计算复杂度,同时利用窗口移动机制增强模型表达能力。不同于文献[12]中窗口移动机制,Sun 等人[13]2022 年提出分层时间窗口注意力机制(hierarchical temporal window attention,HTWA),在多个注意力层中逐层扩大注意力窗口大小,有效捕获远程依赖性的同时,也能够获得不同感受野下更加丰富的多尺度关联特征。考虑到现有 Transformer 模型并未对动作主体对象信息进行建模,Herzig 等人[14]在 2022 年提出对象区域视频 Transformer(object-region video Transformers,ORViT)网络特征优化模块,可以将视频中对象边界框信息融合到网络 Transformer 层中,从而提升模型的性能,该方法需要借助外部目标检测器提取出视频数据中对象边界框信息。

基于 3DCNN 和使用 Transformer 的网络虽然在视频识别领域成为主流,但这些方法参数量大,从头训练对图形处理器(graphics processing unit,GPU)性能和显存要求高,相比之下,2D 卷积计算成本低,同时也可以通过添加高效时间建模模块实现捕获时序依赖关系的功能。例如,Lin 等人[15]在 2019 年将时间移位操作融入 2D 卷积网络中构成时间移位(temporal shift module,TSM)网络,以几乎零计算成本的额外操作使得网络获得时序关联建模能力并且保持较低复杂度,为基于 2D 卷积的网络模型在动作识别领域应用提供新思路,但是 TSM 只对局部时间关联信息建模,而忽略了较为重要的远程时间依赖关系。不同于 TSM 中时间移位操作,Hu 等人[16]在挤压激励网络(squeeze-and-excitation networks,SE-Net)中提到,多尺度池化可以提取不同感受野下依赖信息。在这个结论启发下,同时为建模多粒度时空信息,Chen 等人[17]在 2023 年提出动作粒度金

字塔模块(action granularity pyramid network,AGPN)依次从上到下减小感受野,使提取到的时空信息更加精细化。Tan 等人[18]在 2021 年构建选择性依赖聚合(selective dependency aggregation,SDA)模块,该模块使用多尺度池化进行特征压缩来建模远程和短程时空依赖关系,并且对多种依赖特征进行选择性融合。但是,只使用简单池化操作提取远距离关联信息不可避免地会有信息丢失,具有一定局限性。为解决这一问题,同时针对现有细粒度动作识别模型提取特征尺度单一、对远程时间依赖信息不敏感的缺点,本文提出时空多尺度关联特征融合 2D 卷积网络模型。为增加提取特征丰富性,利用多尺度池化压缩然后激发特征,得到包含不同感受野信息的多尺度特征表示;为提升模型捕获远程时序信息的能力同时尽可能减少额外引入参数量,使用窗口自注意力并且只在时间维度计算自注意力权重来对远程依赖信息建模;为进一步提升特征丰富性并去除冗余,特征融合模块为重要性较强特征赋予更大权重实现自适应加权特征融合,得到更加高效的特征表示。另外,本文选用 TSM ResNet-50 作为动作识别主干网络,既降低模型复杂度,又可保持较强特征描述能力。

1　时空多尺度关联特征融合的二维卷积网络模型

1.1　时空多尺度关联特征融合网络

本文提出的时空多尺度关联特征融合网络,用于捕获视频数据在时间维度和空间维度上的多个尺度依赖关系,并自适应地融合这些关联特征,其目标是获得细粒度视频数据更精细的、高效的特征表示,以提高信息利用效率。该模型由 3 个模块组成,即时间依赖建模模块(temporal dependency modeling,TDM)、空间依赖建模模块(spatial dependency modeling,SDM)和特征融合模块(feature fusion)。模型整体框架如图 1 所示。首先,使用 1×1×1 卷积对来自残差块卷积层 Conv2 的特征进行通道缩减,以减少参数数量。然后,空间依赖建模模块在全局和局部尺度上压缩、激发特征,从而进行短程和远程的空间依赖建模。接下来,时间依赖建模模块将输入特征划分为多个窗口,在每个窗口中计算注意力权重并加权输出。随后,自适应特征融合模块利用查询结构化自注意力机制,按照不同权重加权融合所

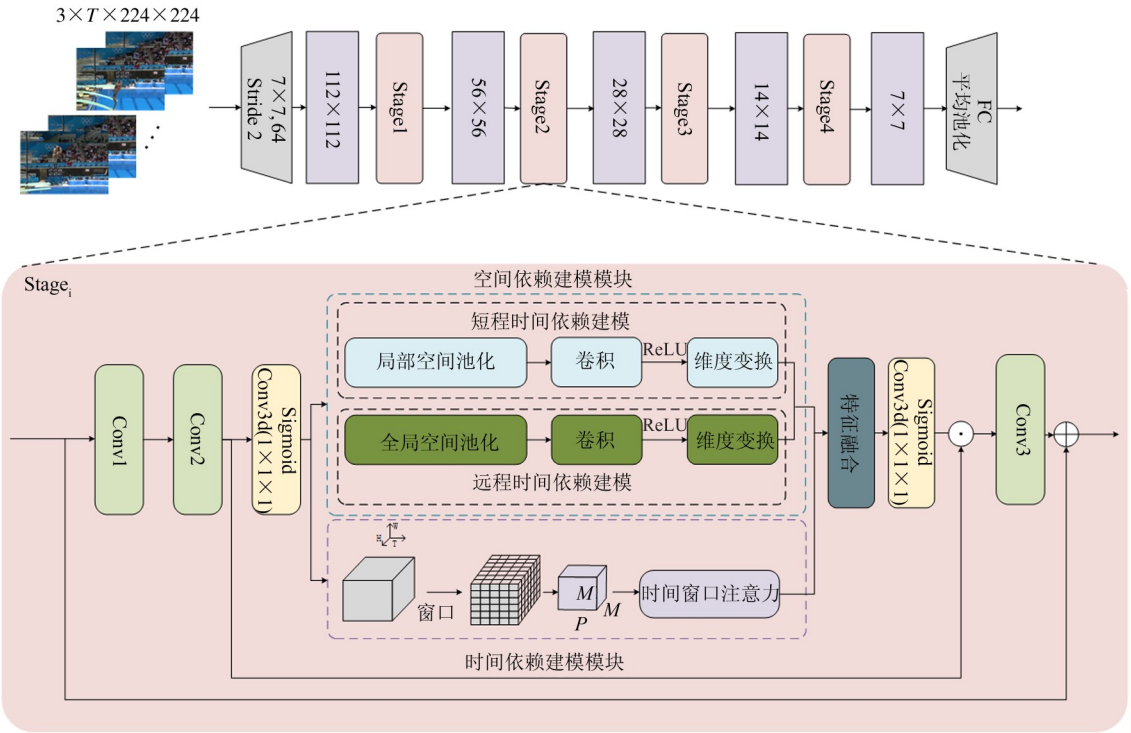

图 1　多尺度时空关联特征融合网络

获得的时空关联特征。最后,通过 1×1×1 卷积层恢复特征通道数,并将输出特征传递给深层网络以进行后续训练和分类。

1.2 基于 ResNet-50 的主干网络

ResNet[19]通过引入残差连接,成功解决由于网络过深而产生的梯度消失和梯度爆炸问题,从而实现更有效的训练过程,因此具备良好的识别精度,同时相比于其他深层网络参数较少,已经成为计算机视觉领域中被广泛关注的模型,因此本文将提出的时空多尺度关联特征融合模块(multi-scale spatio-temporal correlation feature fusion,MSTF)插入到现有 Res-Net 变体网络 TSM ResNet-50 中来构成本文的时空多尺度关联特征融合网络(MSTF-Net)。

1.2.1 ResNet-50 网络

ResNet-50 的主要部分由 4 个网络模块构成,每个模块被称为一个阶段(Stage)。在每个 Stage 中,首先用 1×1 卷积减小输入特征图通道数,达到降维目的,有利于特征提取和后续网络的学习;接着使用 3×3 卷积来细化特征;然后使用多个 1×1 卷积实现特征升维,高维特征可以提供更多模式类别信息,有利于分类;最后输出特征与输入特征进行残差连接。

1.2.2 TSM ResNet-50

由于 ResNet-50 仅在视频数据的单帧上执行特征提取,未能捕获时间维度上的关联信息,为解决这个问题,时间通道位移模块 TSM[15]通过进行通用且几乎无计算成本的时间通道移位操作,在时间维度上将部分通道进行移动,从而实现视频中相邻帧之间信息交互。这种方法在一定程度上实现低成本短程时间依赖建模,满足模型在实际应用中对低延迟实时分类的需求。将 TSM 模块直接插入卷积层或残差块之前会破坏网络在空间建模方面的能力,较为理想的做法是将 TSM 模块嵌入到残差分支中,如图 2 所示。

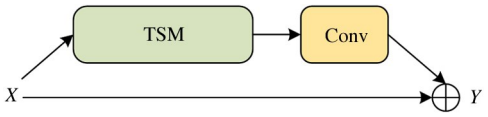

图 2 TSM 操作插入位置示意图

由于 ReNet-50 提取空间特征尺度较为单一并且未对数据时间维度信息进行建模,同时 TSM 模块在时间维度上进行通道移位仅仅促进了相邻帧之间的短程信息交流,未提取丰富的远程时间依赖关系,因此本研究将提出的 MSTF 模块插入 TSM ResNet-50 残差块中构成 MSTF-Net 来缓解上述挑战。具体来说 MSTF 空间依赖建模模块建模多尺度空间关联信息;时间依赖建模模块通过引入自注意力机制捕获远程时间关联信息以此来弥补 TSM 的不足。网络的整体结构以及不同网络结构之间的对比如表 1 所示,表中方括号内的信息表示残差块内部的卷积层结构,每个阶段由多个这样的结构堆叠而成,该数量用方括号外乘积表示。

表 1 ResNet-50 网络结构以及本文模型网络结构

网络层	输出特征大小	ResNet-50	MSTF ResNet-50
Conv1	112×112	7×7,64,stride2	
		3×3 maxpool,stride2	
$Stage_1$	56×56	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}TSM\\1\times1,64\\3\times3,64\\MSTF\\1\times1,256\end{bmatrix}\times3$
$Stage_2$	28×28	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}TSM\\1\times1,128\\3\times3,128\\MSTF\\1\times1,512\end{bmatrix}\times4$
$Stage_3$	14×14	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}TSM\\1\times1,256\\3\times3,256\\MSTF\\1\times1,1024\end{bmatrix}\times6$
$Stage_4$	7×7	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}TSM\\1\times1,512\\3\times3,512\\MSTF\\1\times1,2048\end{bmatrix}\times3$
Classifier	Average pool , FC, softmax		

1.3 空间依赖建模

为更好地捕获多尺度细粒度空间关联信息,便于后续网络进行信息融合和分类,本文构建空间依

赖建模(SDM)模块。具体分为 2 个部分:远程空间依赖建模(short-range spatial dependency)和短程空间依赖建模(long-range spatial dependency)。

给定网络中间层特征 $X \in R^{T\times H\times W\times C}$,其中 C 表示特征通道数,T 表示时间维度,H 和 W 分别表示视频采样帧高度和宽度。首先使用 $1\times1\times1$ 卷积对特征在通道维度上降维,得到降维特征 X',以减小计算量。然后,空间依赖建模(SDM)模块以降维特征 X' 为输入,输出包含空间关联信息的特征 R^S,表示为

$$R^S = \mathrm{SDM}(X') \tag{1}$$

其中 $R^S \in R^{T\times H\times W\times C}$。

具体来说,采用特征压缩、特征激发的级联机制来提取空间关联特征,如图 3 所示。SDM 首先使用平均池化对降维特征 X' 进行压缩,得到压缩特征 A,表示为

$$A = \mathrm{Poolavg}(X';\ W^{Pool}_{pt,ph,pw}) \tag{2}$$

其中,Poolavg 为平均池化,$W^{Pool}_{pt,ph,pw}$ 代表进行的操作为池化操作,pt、ph、pw 分别代表所采用池化核的维度大小。压缩特征 A 包含不同感受野下的空间信息,通过改变池化核大小可以提取不同尺度的压缩信息。然后,使用相应卷积和激活函数对压缩特征 A 进行特征激发得到空间关联特征 R^S,用式(3)表示。

$$R^S = \mathrm{ReLU}(\mathrm{Conv3d}(A; W^{Conv}_{ct,ch,cw})) \tag{3}$$

其中 Conv3d 代表 3D 卷积运算。

对于视频数据,远程空间关联性表示大感受野下宏观内容信息,而短程空间关联性则包含局部范围内细节内容信息。这 2 种关联信息对于细粒度动作识别都比较重要。

为捕获远程空间关联信息(LS)将池化核尺度设置为 $W^{pool}_{1,H,W}$,然后通过平均池化操作对 X' 进行特征压缩得到远程压缩特征 A^{LS},用式(4)表示。

$$A^{LS} = \mathrm{Poolavg}(X';\ W^{Pool}_{1,H,W}) \tag{4}$$

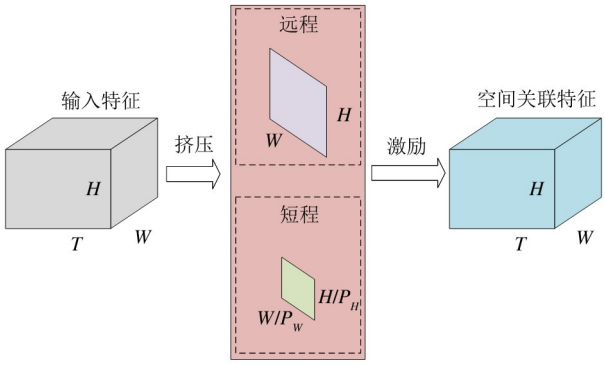

图 3　特征压缩和特征激发机制示意图

远程压缩特征 A^{LS} 中包含长宽维度上跨空间压缩的全局信息。最后通过卷积层和 ReLU 激活函数对 A^{LS} 进行特征激发得到远程空间关联特征 R^{LS},表示为

$$R^{LS} = \mathrm{ReLu}(\mathrm{Conv3d}(A^{LS}; W^{Conv}_{1,3,3})) \tag{5}$$

与远程依赖建模类似,短程依赖建模是通过将池化核尺寸设置在较小范围内实现的。参考文献[18]对局部池化核的设置,本文使用 $W^{pool}_{1,1,2}$ 作为池化核来进行特征压缩,然后通过卷积层 $W^{Conv}_{3,1,1}$ 和 ReLU 激活函数对压缩特征 A^{SS} 进行特征激发,得到短程空间关联特征 R^{SS},这一过程如式(6)所示。

$$\begin{aligned} A^{SS} &= \mathrm{Poolavg}(X;\ W^{Pool}_{1,2,2}) \\ R^{SS} &= \mathrm{ReLU}(\mathrm{Conv3d}(A^{SS};\ W^{Conv}_{3,1,1})) \end{aligned} \tag{6}$$

1.4　时间依赖建模

为提取特征时间维度上远程依赖关系,引入时间依赖建模模块。该模块使用自注意力机制构建远程依赖关系,但直接将所有视频采样帧输入到 Transformer 网络中进行全局自注意力学习会导致计算成本较高的问题,因此,本文引入时间窗口自注意力机制(temporal window self attention, Temporal W-SA),它仅在窗口内对输入数据进行自注意力操作,并且只需要关注时间维度上依赖关系。时间窗口自注意力机制计算流程图如图 4 所示。

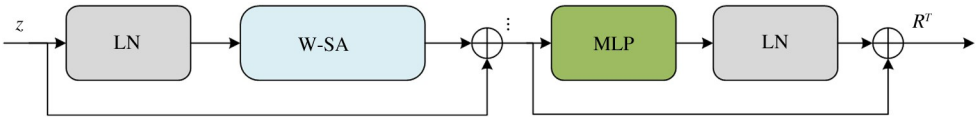

图 4　时间窗口自注意力计算流程图

假设每个通道特征维度为 $T \times H \times W$,使用一个维度为 $P \times M \times M$ 的 3D 窗口以非重叠方式将输入特征划分,输入特征被划分为 $\frac{T}{P} \times \frac{H}{M} \times \frac{W}{M}$ 个不重叠窗口。例如,对于维度为 $8 \times 8 \times 8$ 的输入特征,如果采用 $2 \times 2 \times 2$ 窗口进行划分,最终划分得到 64 个窗口。然后,在每个 3D 窗口内进行自注意力操作。

时间关联建模模块主要关注时序关系建模,因此时间依赖建模模块只在每个窗口时间维度上对特征进行自注意力操作。输入特征被划分为 k 个窗口,首先每个窗口中特征在空间维度分割成固定大小非重叠块(patch),通过图像块编码(patch embedding)和位置编码(position embedding)[12]得到带有位置信息的向量 z;接下来,归一化层(layer normalization,LN)对向量 z 进行归一化处理以提高模型稳定性[20];紧接着,在时间维度上对 LN 层输出进行自注意力操作,并将其与输入 z 进行残差连接,得到中间特征向量 $\hat{z}$。这一过程表示为

$$\hat{z} = \text{Temporal W-SA}(\text{LN}(z)) + z \tag{7}$$

然后,通过由多层感知机(multilayer perceptron,MLP)和 LN 组成的前向传播模块对中间向量 $\hat{z}$ 作进一步特征提取,以增强特征表达能力,并与 $\hat{z}$ 进行残差连接得到输出特征 a_i。最后,将各个窗口输出拼接在一起,得到远程时间关联特征 R^T。该步骤表示为

$$\begin{aligned} a_i &= \text{LN}(\text{MLP}(\hat{z})) + \hat{z},\ i \in [1,2,\cdots,k] \\ R^T &= \text{Concat}(a_1, a_2, \cdots, a_k) \end{aligned} \tag{8}$$

其中,a_i 为多层感知机的输出特征,Concat 代表拼接操作,$i \in [1,2,\cdots,k]$,其中 k 表示划分的窗口数量。

由于本文使用固定且不重叠的时间窗口,故而缺乏跨窗口信息连接,这可能会限制模型特征提取能力。为此,在实验部分本文对比不同 P 值窗口下模型性能,最终得出性能最优时窗口时间长度。

1.5 特征融合

为得到更加高级、紧凑的特征表示,以便于后续分类,引入特征融合模块。由于视频内容之间关联性较为复杂,并且在不同类型视频中,语义信息的类别依赖于不同关联信息,因此这些关联特征对动作识别的贡献可能是不平等的。本文充分考虑视频内容中各种关联信息多样性,为保证融合得到最合理紧凑的时空关联特征以促使模型更好地理解视频动作语义信息,引入自适应权重特征融合(adaptive fusion,ADF)模块,利用自注意力机制动态生成不同大小的注意力权重以此来强调较为重要的特征,模块的框图如图 5(a)所示。

依赖建模模块输出一组时空关联特征 $\{R_1, R_2, \cdots, R_M\}$,特征融合模块对多个特征进行融合。ADF 模块受到查询结构化自注意力[21]的启发,额外引入一个可学习查询向量 $\boldsymbol{Q} \in R^{1\times C}$ 来辅助计算相应自注意力权重。首先将 $\{R_1, R_2, \cdots, R_M\}$ 简单拼接后,得到向量 $\boldsymbol{V} \in R^{M\times T\times H\times W\times C}$,然后对向量 $\boldsymbol{V}$ 进行平均池化,向量 $\boldsymbol{V}$ 在时间和空间上被压缩,进而得到 $\boldsymbol{K} \in R^{M\times C}$,最后计算得到得到注意力权重,计算过程表示为

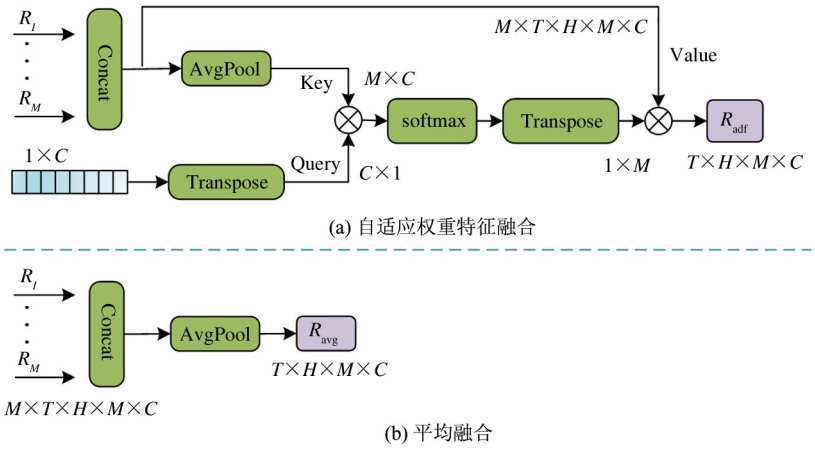

图 5 2 种特征融合方式

$$\mathrm{Attention}(\boldsymbol{Q}, k) = \mathrm{softmax}(\boldsymbol{Q} \times \boldsymbol{K}^T) \tag{9}$$

最终的多尺度时空关联特征 $R_{\mathrm{adf}} \in R^{T\times H\times W\times C}$ 计算如下：

$$R_{\mathrm{adf}} = \mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}) \times \boldsymbol{V} \tag{10}$$

为进一步验证本文 ADF 模块有效性，引入较为传统的平均融合（average fusion，AVF）模块来与 ADF 进行对比。AVF 不考虑关联特征之间重要性区别，为不同关联特征赋予相同权重来进行特征融合，如图 5（b）所示。首先将 $\{R_1, R_2, \cdots, R_M\}$ 拼接起来，然后使用平均池化最终得到关联特征 $R_{\mathrm{avf}} \in R^{T\times H\times W\times C}$。计算公式表示为

$$R_{\mathrm{avf}} = \frac{1}{M}\sum_{i=1}^{M} R_i \tag{11}$$

其中 $i \in [1, 2, \cdots, M]$，M 为特征数量，$\{R_1, R_2, \cdots, R_M\}$ 为依赖建模模块输出的一组特征。为验证所引入 ADF 模块的优势，同时确定最终应采用的特征融合方式，本文在实验部分对使用 2 种融合方式的模型识别精度进行对比。

2　实验与分析

本文提出的时空多尺度关联特征融合 2D 卷积网络在 PyTorch 深度学习框架下设计并实现。使用 4 个 Tesla K80 GPU 训练，并且对于输入 8 帧的模型，批大小（batch size）设置为 32，初始学习率设置为 0.010；对于输入 16 帧的模型，Batch size 设置为 16，初始学习率设置为 0.007 5；训练 50 个迭代（epoch）并且在第 20、40 个 epoch 时学习率衰减为原来的 0.1。

2.1　数据集

如图 6 所示，Diving48 数据集是通过剪辑和切分跳水比赛视频获得的。与常用的粗粒度数据集（如 UCF101、HMDB51）相比，Diving48 数据集对各种细微的动作进行了定义分类，粗粒度的背景信息对分类贡献不大，因此该数据集可以用于评估模型对远程时序关联信息的捕获能力。Diving48 数据集包含 48 个类别，第 1 个版本的数据集一共有 18×10^3 视频，但存在较多错误标注和一些剪辑不准确的视频，于是官方于 2020 年发布第 2 版本，并在该版本进行删减。删减后的 Diving48-V2 共包含 17×10^3 视频，随机划分 15×10^3 用作训练集，2×10^3 用作测试集，并且视频的长度、帧率不等。本文实验都是在 Diving48-V2 以及官方提供的训练集和测试集下进行的。

图 6　Diving48 细粒度动作数据集展示

如图 7 所示，Something-somethingV1 数据集是一个大规模的细粒度动作识别数据集，包含 108 499 个 2～6 s 有标注视频片段，174 个动作类别，例如：拿起某物、打开某物等。此数据集强调人类与物体之间交互关系，要求模型能够提取不同帧之间细粒度时间关联信息。

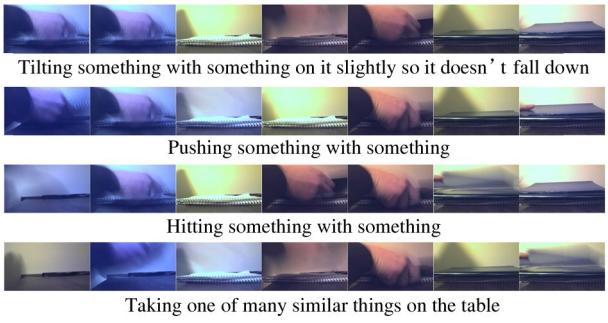

图 7　Something-somethingV1 数据集展示

2.2　消融实验

为研究本文模型各个部分的有效性，本节在 Diving48-V2 数据集上设置 3 组消融实验，分别为时间窗口长度的影响、依赖建模模块不同部分的作用以及特征融合方式对识别精度的影响。

2.2.1　不同时间窗口长度的影响

当使用固定注意力窗口划分输入特征时,不同窗口间视频帧依赖关系没有被模型识别。为研究这些丢失信息对识别精度的影响,并进一步确定最优的窗口大小,本文分别在 P 等于2、4、8 和16 时进行实验,对比研究时间窗口大小对模型精度的影响,实验结果如图 8 所示。

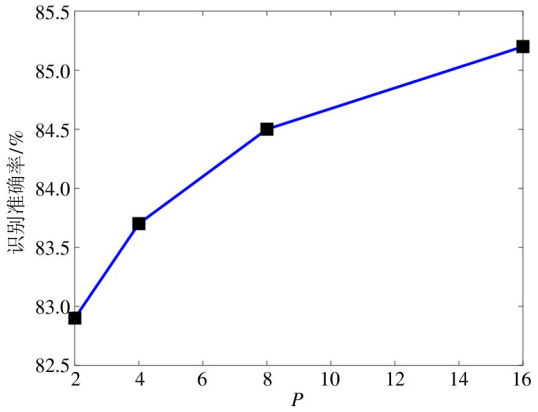

图 8　窗口时间长度 P 对识别精度的影响

从图 8 可以看到,随着 3D 窗口的时间长度 P 的增加,模型的识别准确率不断增加;当 P 增加到与输入帧数 16 相等的时候,模型性能达到最优。实验结果表明,较小的时间注意力窗口不利于远程时间依赖建模。在增加时间窗口长度后,虽然在相同条件下模型训练时间有小幅增加,但模型分类准确率却有较为显著的提升。

2.2.2　模块不同部分的作用

为验证模型各个部分有效性,本节对模型中时间依赖建模模块和空间依赖建模模块的作用进行消融研究,模型依赖建模模块各部分输入和输出特征图维度均相同,因此每部分可以单独使用。对于空间关联建模模块,将池化核和卷积核分别设置为 $\{1\times H\times W, 1\times 3\times 3\}$ 来进行远程空间依赖建模,将池化核和卷积核大小分别设置为 $\{1\times 2\times 2, 3\times 1\times 1\}$ 来进行短程空间依赖建模;对于时间关联建模模块,将时间窗口长度分别设置为 4、8、16,来测试时间关联建模模块的时序建模能力;特征融合方式则采用平均融合。实验结果如表 2 所示,当单独使用空间建模模块时,模型性能提升 0.2%;单独使用时间依赖建模模块时,增大窗口的时间长度依然可以不断提升模型的精度直到窗口长度增加到与输入帧数相等时精度达到最高,可以使模型性能提升 1.3% ~ 2.5%;同时使用 2 种依赖建模模块则会使性能最多提升 3%。总体而言,无论为网络增加哪种依赖建模模块,都会使原始 TSM 网络识别准确率得到提高。实验结果表明,本文提出的时空依赖建模模块各个部分都是有效的,能够提取多尺度时空关联特征,并且提高细粒度动作识别准确率。

表 2　SDM 模块和 TDM 模块作用对比实验

依赖	描述	窗口	准确率/%
空间	远程空间(LS) + 短程空间(SS)	-	82.4
时间	Short window	4×3×3	83.5
	Medium window	8×3×3	84.1
	Long window	16×3×3	84.7
	LS + SS + Long window		85.2
	TSM		82.2

另外,实验结果显示,时间依赖建模模块对性能提升的贡献似乎大于空间依赖建模模块,本文推测导致这样结果的原因如下:首先,基于 2D 卷积的主干网络已经具有较强空间建模能力,继续加强空间关联提取能力对模型识别准确率有一定提升但不显著;其次,时间依赖建模模块弥补 TSM 模块和 ResNet-50 主干网络远程时间依赖建模能力不足的缺点,从而带来显著性能提升。

2.2.3　特征融合方式对识别准确率的影响

不同关联特征对视频动作识别有不同贡献率,为验证推断,同时确定适当的特征融合方式,本节对比研究模型分别在 AVF 方式和 ADF 特征融合方式下识别精度。分别在窗口时间长度 P 为 4、8、16 时进行实验,实验结果如表 3 所示。在不同时间窗口长度下使用 ADF 特征融合方式的模型准确率均高于使用 AVF 融合方式的模型准确率。这说明不同视频样本的识别需要强调不同关联特征,AVF 方式为特征赋予相同权重则忽略了这一点,相比之下 ADF 特征融合模块能够使用自注意力机制为不同特征动态赋予不同权重来强调有利于识别动作类的特征实现更加有效的特征融合。实验结果表明,

ADF 融合方式更加有效,因此本文确定 ADF 为最终特征融合方式。

表 3　不同特征融合方式对比实验

特征融合方式	窗口长度 P	准确率/%
AVF	4	83.7
ADF		84.3
AVF	8	84.5
ADF		85.7
AVF	16	85.2
ADF		86.0

2.3　与其他方法对比

为进一步验证本文方法有效性,本节设置了模型与一些主流方法在识别准确率上的对比实验,这些方法包括双流方法、基于 3D 卷积的方法和基于 Transformer 的方法。另外,本节分别在输入帧数为 8 帧和 16 帧情况下进行实验,目的是验证算法在远程时间依赖建模上的优势。

2.3.1　在 Diving48 数据集上性能对比

首先在每个视频中均匀采样 8 帧,然后将每个帧裁剪为 224 × 224 的大小进行实验,实验结果如表 4所示。数据显示,本模型将主干网络 TSM 的识别准确率提升 1.9%,达到 79.5%。与同样输入 8 帧的方法对比,虽然只使用视频帧(RGB)信息一种模态作为输入,但是本文模型能够直接从 RGB 信息中同时建模时间和空间信息。然而,双流网络 TSN 靠同时输入光流和 RGB 数据来分别捕获运动信息和空间信息,对多尺度的时空关联不敏感,同时时空信息之间的交融不充分,因此本文模型识别准确率超过 TSN 模型 7.1%;虽然基于 2D 卷积网络,但是本模型对多尺度时空信息,尤其是远程时间关联信息更为敏感,相比之下,基于 3D 卷积的网络,如表中的 C3D、S3D、SlowFast 等模型由于受卷积核尺寸限制,不利于捕获远程关联信息,因此其识别精度均落后于本文模型(5.3% ~28.9%)。实验结果表明,本文提出的模块可以使 2D 卷积网络具有像 3D 卷积网络一样的时序依赖关系建模能力,但是 8 帧的模型性能比不过最近提出基于 Transformer 的方法,接下来将输入帧增加到 16 帧进行实验。

表 4　在 Diving48 数据集上与其他方法对比

方法	主干网络	输入帧数	模态	识别准确率/%
ST-S3D[9]	S3D	8	RGB	50.6
C3D[5]	3DResNet-50	8	RGB	73.4
GST[22]	ResNet-50	8	RGB	74.2
TSN[4]	ResNet-50	8	RGB + 光流	72.4
TSM[15]	ResNet-50	8	RGB	77.6
本文方法	ResNet-50	8	RGB	79.5
SlowFast[7]	ResNet101	128	RGB	77.6
LT-S3D[9]	S3D	all frames	RGB	80.4
Timesformer[11]	Timesformer	16	RGB	74.9
Timesformer-L[11]	Timesformer	96	RGB	81.0
TQN[9]	S3D + Feature Bank	all frames	RGB	81.8
TSM[15]	ResNet-50	16	RGB	82.2
RSANet[23]	ResNet-50	16	RGB	84.2
ORViT[14]	Timesformer	32	RGB	88.0
本文方法	ResNet-50	16	RGB	86.0

将输入帧增加到 16 帧后,受到 GPU 显存限制,batch size 减小到 16,同时学习率对应减小到 0.075。当将输入帧数增加到 16 帧时,识别精度从 79.5% 提升到了 86.0%,本文模型将 TSM 精度提升了 3.8%。提升幅度较大,这表明本模型在识别帧数较多的视频时有一定优势,更适用于远程时间依赖建

模。与同样输入 16 帧或者更多帧的方法对比,模型识别精度超过需要密集输入帧的长时间可分离 3D 卷积网络(long-time separable 3D convolutions, LT-S3D),和需要输入 128 帧的 SlowFast 网络,这再次证明本模型相比于传统 3D 卷积网络具有更好的远程依赖建模能力。与最近的基于 Transformer 的方法相比,本文方法识别精度高于输入 16 帧的 Timesformer 和输入所有帧的 TQN,展现出较强时空关联信息提取能力。但是最先进的 ORViT 识别精度高于本文方法2.0%,分析原因如下:首先,本文模型基于 ResNet-50 主干网络,ORViT 基于 Timesformer 主干网络,参数量远远大于 ResNet-50(121.4×10^6 VS 25.6×10^6);其次,ORViT 通过将视频中动作主体对象边界框融入到模型训练过程中实现的以对象为中心的动作识别,需要外界先验知识辅助学习,而本文模型只需要输入视频帧信息用于训练;最后,ORViT 需要 32 帧输入,而本模型为 16 帧输入计算量更小。综上所述,本文模型相比 ORViT 在以上几个方面具有一定优势。

2.3.2 在 Something-something V1 数据集上的性能对比

为了进一步验证本方法的普适性,本文在 Something-something V1 数据集上进行输入帧为 8 帧的实验,实验结果如表 5 所示。在该数据集上,本文算法将 TSM 模型识别精度提升了 1.3%,加强了主干网络时空关联建模能力。与经典 3D 网络双流膨胀 3D 卷积网络(I3D)、及 I3D 的 2 个变体(I3D 加非局部 Non-Local 算子组成的 NLI3D 和结合图卷积神经网络(graph convolutional networks, GCN)组成的 NLI3D + GCN),以及 3D 卷积改进的虚拟视图合成(virtual view synthesis, VVS)网络对比,本文模型识别准确率均有一定的优势,这表明本文提出模块能够使 2D 网络具有 3D 卷积网络的时空建模能力。相比于经典双流方法 TSN,本文方法识别精度优势较大,原因可能是 TSN 只是通过池化方式进行时间建模,并且简单融合 RGB 流和光流的分类结果,对远程复杂时间依赖关系建模不足。与一些基于 2D 卷积网络的方法相比本文方法也具有一定优势,其中,TAM(temporal adaptive module)通过时间自适应模块捕获局部和长期时间关系来使 2D 卷积网络实现时间建模,时空分离建模网络(space-time separate model, STS)通过将特征通道进行分组并分别对每组中的特征进行不同类型的卷积操作来实现时空建模。本文算法相比于 STS 和 TAM,在识别精度上具有一定优势,原因可能是本文使用的自注意力机制具有更加优越的远程依赖关系捕获能力,使模型能够更好地利用视频中的细粒度信息。

表 5 在 Something-something V1 数据集上与其他方法对比

方法	主干网络	输入帧数	模态	准确率/%
I3D[6]	3DResNet-50	32	RGB	41.6
NLI3D[8]	3DResNet-50	32	RGB	44.4
NLI3D + GCN[24]	3DResNet-50	32	RBG	46.1
VVS[25]	I3D + 3DResNet-50	-	RGB	46.1
TSN[4]	ResNet-50	8	RGB + 光流	19.7
ECO[26]	ResNet-18	8	RGB	39.6
STS[27]	ResNet-50	16	RGB	46.0
TAM[28]	ResNet-50	8	RGB	46.5
TSM[15]	ResNet-50	8	RGB	45.6
本文方法	ResNet-50	8	RGB	46.9

3 结 论

为解决传统 2D 卷积网络提取时空特征尺度单一、以及对细粒度动作识别数据集中远程时间关联信息识别率不足的问题,提出时空多尺度关联特征融合的 2D 卷积网络细粒度动作识别模型,提高 2D

网络对视频数据时空关联信息的建模能力。在 2 个公开细粒度数据集上进行实验,结果表明:本文模型的时间窗口自注意力机制可以有效地提取视频数据中丰富的远程时间关联信息;多尺度池化能够实现数据中细粒度远程、短程空间依赖建模,得到更加丰富的特征表示;特征融合模块可以自适应地使用更大的权重来强调重要特征,从而实现特征自适应融合。此外,本文模型基于 2D 卷积,并且只使用视频帧信息作为输入,网络结构简单且易于训练。提出的方法在公开数据集上具有良好识别精度,并且可以进一步提高主干网络性能,对细粒度动作识别算法研究具有一定借鉴意义。

参考文献

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

[2] YUE-HEI NG J, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015:4694-4702.

[3] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:1933-1941.

[4] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016:20-36.

[5] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 4489-4497.

[6] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 6299-6308.

[7] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019:6202-6211.

[8] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:7794-7803.

[9] ZHANG C, GUPTA A, ZISSERMAN A. Temporal query networks for fine-grained video understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online: IEEE, 2021: 4486-4496.

[10] BERA A, NASIPURI M, KREJCAR O, et al. Fine-grained sports, yoga, and dance postures recognition: a benchmark analysis[J]. IEEE Transactions on Instrumentation and Measurement, 2023,72:1-13.

[11] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding? [C]//International Conference on Machine Learning. Vienna, Austria: PMLR, 2021:813-824.

[12] LIU Z, NING J, CAO Y, et al. Video swin transformer [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022:3202-3211.

[13] SUN Y, XUE H, SONG R, et al. Long-form video-language pre-training with multimodal temporal contrastive learning[J]. Advances in Neural Information Processing Systems, 2022,35:38032-38045.

[14] HERZIG R, BEN-AVRAHAM E, MANGALAM K, et al. Object-region video transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 3148-3159.

[15] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019:7083-7093.

[16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:7132-7141.

[17] CHEN Y, GE H, LIU Y, et al. AGPN: action granularity pyramid network for video action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023,33(8):3912-3923.

[18] TAN Y, HAO Y, HE X, et al. Selective dependency aggregation for action classification[C]//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM, 2021:592-601.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:770-778.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. (2017-06-12) [2023-1128].

https://arxiv.org/pdf/1706.03762v1.pdf.

[21] HAO Y, ZHANG H, NGO C W, et al. Compact bilinear augmented query structured attention for sport highlights classification[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM, 2020:628-636.

[22] LUO C, YUILLE A L. Grouped spatial-temporal aggregation for efficient action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019:5512-5521.

[23] KIM M, KWON H, WANG C, et al. Relational self-attention: what's missing in attention for video understanding[J]. Advances in Neural Information Processing Systems, 2021,34:8046-8059.

[24] WANG X, GUPTA A. Videos as space-time region graphs[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 399-417.

[25] PENG G, LI Y L, ZHU H, et al. VVS: action recognition with virtual view synthesis[C]//2021 IEEE International Conference on Image Processing. Anchorage, USA: IEEE, 2021:384-388.

[26] ZOLFAGHARI M, SINGH K, BROX T. ECO: efficient convolutional network for online video understanding[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018:695-712.

[27] CAO P, WANG S, WU J, et al. Space-time separate modeling for efficient video classification[C]// The 2nd International Conference on Computer Vision and Data Mining. Changsha, China: IOP Publishing, 2021:1-6.

[28] LIU Z, WANG L, WU W, et al. Tam: temporal adaptive module for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021:13708-13718.

Fine-grained 2D convolutional network model for action recognition based on spatio-temporal multi-scale correlation feature fusion

HU Zhengping[*,**], WANG Xinyu[*], DONG Jiawei[*], ZHAO Yanshuang[*], LIU Yang[*]

([*]School of Information and Engineering, Yanshan University, Qinhuangdao 066004)

([**]Hebei Key Laboratory of Information Transmission and Signal Processing, Yanshan University, Qinhuangdao 066004)

Abstract

In order to solve the problems of traditional 2-dimensional (2D) convolutional network extracting spatiotemporal features at a single scale and insufficient utilization of long-range temporal correlation information between frames in fine-grained action data sets, this paper proposes a fine-grained 2D convolutional network model for action recognition based on spatio-temporal multi-scale correlation feature fusion model. First, in order to model the multi-scale spatial correlation of videos to enhance the spatial representation ability of fine-grained video data, the model uses a multi-scale 'feature squeeze and feature excitation' method to make the spatial features extracted by the network more abundant and effective. Then, in order to fully utilize the motion information in the time dimension of fine-grained video data, a temporal window self attention mechanism is introduced, and the powerful long-range dependency modeling ability of Transformer is utilized to only perform self attention operations in the time dimension, modeling long-range time dependencies at a lower computational cost. Finally, considering that the extracted spatio-temporal features contribute unevenly to different types of action classification, an adaptive feature fusion module is introduced to dynamically assign different weights to features to achieve adaptive feature fusion. The model's Top-1 accuracy on the two fine-grained action recognition data sets Diving48 and Something-somethingV1 reached 86.0% and 46.9% respectively, which improved the Top-1 accuracy of the original backbone network by 3.8% and 1.3% respectively. Experimental results show that when only using video frame information as input, this model achieves recognition accuracy comparable to existing algorithms based on Transformer and 3-dimensional convolutional neural network (3D CNN).

Key words: fine-grained action recognition, multi-scale spatio-temporal correlation feature, long-range dependency modeling, self-attention mechanism