DOI:10.16136/j.joel.2023.12.0123

基于ConvGRU和注意力特征融合的人体动作识别

程娜娜,张荣芬*,刘宇红,刘 源,刘昕斐,杨 双

(贵州大学 大数据与信息工程学院,贵州 贵阳 550025)

摘要:在动作识别任务中,如何充分学习和利用视频的空间特征和时序特征的相关性,对最终识别结果尤为重要。针对传统动作识别方法忽略时空特征相关性及细小特征,导致识别精度下降的问题,本文提出了一种基于卷积门控循环单元(convolutional GRU,ConvGRU)和注意力特征融合(attentional feature fusion,AFF)的人体动作识别方法。首先,使用Xception网络获取视频帧的空间特征提取网络,并引入时空激励(spatial-temporal excitation,STE)模块和通道激励(channel excitation,CE)模块,获取空间特征的同时加强时序动作的建模能力。此外,将传统的长短时记忆网络(long short term memory,LSTM)网络替换为ConvGRU网络,在提取时序特征的同时,利用卷积进一步挖掘视频帧的空间特征。最后,对输出分类器进行改进,引入基于改进的多尺度通道注意力的特征融合(MCAM-AFF)模块,加强对细小特征的识别能力,提升模型的准确率。实验结果表明:在UCF101数据集和HMDB51数据集上分别达到了95.66%和69.82%的识别准确率。该算法获取了更加完整的时空特征,与当前主流模型相比更具优越性。

关键词:动作识别;注意力机制;ConvGRU;特征融合

中图分类号:TP391.4 **文献标识码**:A **文章编号**:1005-0086(2023)12-1298-09

Human motion recognition based on ConvGRU and attention feature fusion

CHENG Nana, ZHANG Rongfen*, LIU Yuhong, LIU Yuan, LIU Xingfei, YANG Shuang

(College of Big Data and Information Engineering, Guizhou University, Guiyang, Guizhou 550025, China)

Abstract: In the action recognition task, how to fully learn and utilize the correlation between the spatial features and temporal features of the video is particularly important for the final recognition results. Aiming at the problem that the traditional action recognition method ignores the correlation of spatio-temporal features and small features, which leads to the decrease of recognition accuracy, this paper proposes a human action recognition method based on convolutional GRU (ConvGRU) and attentional feature fusion (AFF). Firstly, the Xception network is used to obtain the spatial feature extraction network of video frames, and the spatio-temporal excitation (STE) module and channel excitation (CE) module are introduced to obtain the spatial features and strengthen the modeling ability of temporal actions. In addition, the traditional long short term memory (LSTM) network is replaced by the ConvGRU network, which uses convolution to further mine the spatial features of video frames while extracting temporal features. Finally, the output classifier is improved, and the feature fusion module based on improved multi-scale channel attention is introduced to strengthen the recognition ability of small features and improve the accuracy of the model. The experimental results show that the recognition accuracy of 95.66 % and 69.82 % are achieved on the UCF101 dataset and the HMDB51 dataset, respectively. The algorithm obtains more complete spatio-temporal features and is superior to the current mainstream models.

Key words: action recognition; attention mechanism; ConvGRU; feature fusion

* **E-mail**: rfzhang@gzu.edu.cn

收稿日期:2023-03-21 **修订日期**:2023-06-15

基金项目:贵州省科学技术基金(黔科合基础-ZK[2021]重点001)资助项目

0 引　言

基于视频的人体动作识别是计算机视觉领域的一个重要研究方向,广泛应用于人机交互、智能家居、动作迁移、视频理解与检索等诸多方向[1,2]。与静态图像不同,人体动作视频中的复杂动作通常是与时间相关的,它不仅包含每一帧内的空间信息,还包含一段时间内的时序信息,因此相对于图像分类来说多了一个需要处理的时序维度。

在早期,基于人工特征的改进的密集轨迹(improved dense trajectories,iDT)[3]因其出色的性能,成为行为识别领域的主流方法。然而,由于手工制作的特征,iDT 算法计算成本较高,因此许多研究者将基于卷积神经网络(CNN)的方法应用到动作识别上[4]。SIMONYAN 等[5]提出了双流卷积神经网络(two-stream network),包括利用 RGB 视频帧提取空间特征的空间流和利用堆叠的光流图捕获运动特征的时间流这两个独立的分支,该方法的性能优于以往的手工算法。虽然双流网络在视频中使用了时间信息,但只使用了短期的运动变化,而没有捕捉到视频的远程时间信息。视频序列不同于视频图像,还包含了时序信息,组成了三维向量。因此 TRAN 等[6]使用三维卷积和三维池化设计了 C3D 网络(convoltional 3D),可以同时处理空间和时间维度的信息。

然而,由于 3D 卷积网络参数量庞大,并且只能获取短距离的时空信息和运动信息,难以处理整个视频的长时信息。因此,研究者们将循环神经网络(recurrent neural network,RNN)和长短时记忆网络(long short term memory, LSTM)应用到视频任务中。DONAHUE 等[7]提出了一种长时循环卷积网络(long-term recurrent convolutional networks,LRCN),该网络使用 CNN 提取视频帧的空间特征,并使用 LSTM 对人体动作的时间信息建模,从而实现动作分类。PAN 等[8]介绍了一种跨流选择网络(cross-stream selective networks,CSN),在该 CSN 中,通过 Bi-LSTM 网络为每个 RGB 段选择相关度最高的光流堆栈。JAOUEDI 等[9]使用门控循环单元(gated recurrent unit,GRU),通过高斯混合模型和卡尔曼滤波提取运动数据。此外对于特征提取不够充分的问题,基于多尺度特征提取的算法也被用于动作识别中[10]。

以上方法在动作识别领域取得了一定发展,但目前相关研究还存在以下不足:1) CNN 和 LSTM 只能独立地提取空间特征和时间特征,忽略了时空特征相关性,难以获取更加完整的时空特征;2) 传统的输出分类器仅用全连接层进行最终分类,仅简单提供了特征映射的固定线性聚合,忽略了细小特征,导致识别精度下降。基于此,文章通过在空间特征提取阶段引入注意力模块,分别从时空域和通道域两个维度考虑注意力权重对特征的影响,同时增强空间和时间信息的建模能力,实现对输入特征的自适应调整;在时间特征提取阶段,采用卷积门控循环单元(convolutional GRU,ConvGRU)网络代替传统的 LSTM 网络,提取时间序列信息的同时利用卷积操作进一步提取空间信息;最后运用多尺度特征融合改进输出模型,提高模型识别准确率。

1 模型的构成及其工作原理

1.1 模型总体框架

文章提出了一种基于 ConvGRU 和注意力特征融合(attentional feature fusion,AFF)的人体动作识别算法模型,模型的整体网络框架如图 1 所示。视频预处理部分对每个视频段均匀地采样为视频帧;视频的空间特征的提取方法采用 SC-Xception 网络,该网络考虑到提取空间特征时丢失的时序信息,在基础网络 Xception 中加入了时空注意力和通道注意力机制,提升网络的时序建模能力;视频的时序特征由 ConvGRU 进行提取,利用 ConvGRU 的卷积操作可以同时掌握时间和空间信息;将提取到的特征通过改进的融合多尺度注意力特征的输出分类器,对特征图进行分析处理,得到相应的预测结果。

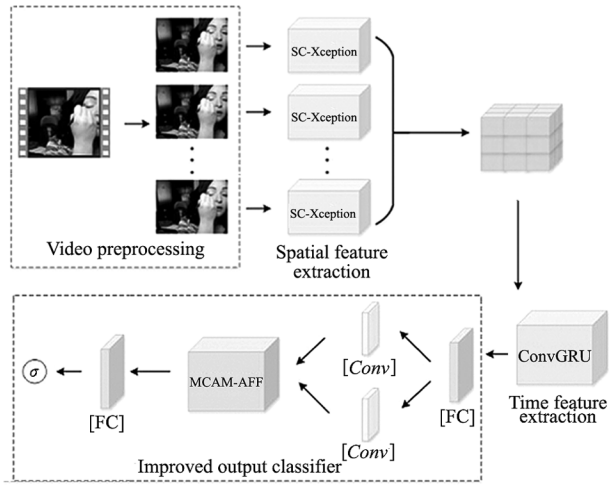

图 1　整体网络架构

Fig. 1　Overall network architecture

1.2 空间特征提取网络

视频帧的空间特征采用轻量级网络 Xception 进行提取。相较于 Inception 系列网络,Xception 采用深度可分离卷积,将通道维度的相关性和空间维度

的相关性分开处理，实现了两个维度的完全解耦，参数量减少的同时提升了模型识别的准确率。

Xception 作为传统的二维卷积，计算成本低，但无法捕捉时序关系，提取空间特征时会忽略掉重要的时序信息。鉴于此，融合时空激励模块（spatial-temporal excitation，STE）和通道激励（channel excitation，CE）模块[11]，将其应用到二维卷积中，在提取空间特征的同时，增强了时序动作的建模能力，进一步补充视频属性的完整性表示。

STE 模块利用三维卷积核，对时间和空间两个维度的信息进行建模，如图 2 所示。式（1）—（4）为 STE 模块的计算式：

$$F^* = R[CAvg(X)], \tag{1}$$

$$F_0 = R[Conv_{3\times3\times3}(F^*)], \tag{2}$$

$$M = \sigma(F_0), \tag{3}$$

$$Y = X \oplus X \odot M, \tag{4}$$

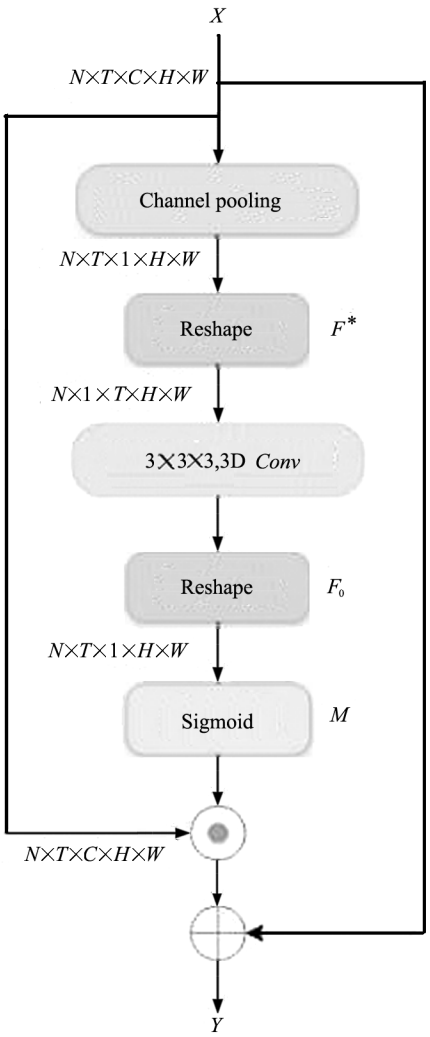

图 2　时空激励模块

Fig. 2　Spatial-temporal excitation module

式中，$CAvg(\cdot)$ 表示通道平均池化，$Conv(\cdot)$ 表示卷积运算，$R(\cdot)$ 表示 Reshape 操作，$\sigma(\cdot)$ 表示 Sigmoid 函数，$\oplus$ 表示向量相加，$\odot$ 表示向量内积，$X \in R^{N\times T\times C\times H\times W}$，$F^* \in R^{N\times1\times T\times H\times W}$，$F$、$F_0 \in R^{N\times T\times1\times H\times W}$。

与传统的三维卷积运算相比，STE 的计算效率更高，因为输入特征是跨通道平均的。输入张量 X 的每个通道都可以从一个细化的特征激励中获取到时空特征的权重信息。

CE 模块的结构与压缩激励（squeeze excitation，SE）模块[12]类似，如图 3 所示。但 CE 在两个卷积核为 1×1 的二维卷积层之间插入了一个一维卷积层来表征通道特征的时间信息。式（5）—（9）为 CE 的计算式：

$$F_r = Conv_{1\times1}[SAvg(X)], \tag{5}$$

$$F_t = R(F_r), \tag{6}$$

$$F_0 = Conv_{1\times1}\{R[Conv_3(F_t)]\}, \tag{7}$$

$$M = \sigma(F_0), \tag{8}$$

$$Y = X \oplus X \odot M, \tag{9}$$

式中，$F_r \in R^{N\times\frac{C}{16}\times T\times1\times1}$，$F_t \in R^{N\times T\times\frac{C}{16}\times1\times1}$，$F_0 \in R^{N\times T\times C\times1\times1}$，$SAvg(\cdot)$ 为空间平均池化。

将 STE 和 CE 模块嵌入到 Xception 网络最后一个残差块的起始点，如图 4 所示，可以得到改进后的 SC-Xception 网络。改进后的 SC-Xception 网络不仅加强了对空间特征的提取能力，同时也具备获取时间序列信息的能力。

1.3　时间特征提取网络

时间特征提取采用的是 ConvGRU 网络[13]，GRU 是 LSTM 的变体形式，相较于 LSTM 少了一个门函数，减少了参数量，保证模型性能的同时提升了训练速度和收敛速度，可以避免在小数据集上难以得到充分训练而导致模型过拟合。GRU 的两个门控机制，即重置门和更新门，能够保存长期序列中的信息，并且不会随着时间流逝或与预测不相关而被清除，可以解决一般的RNN所存在的长期依赖问题。

ConvGRU 使用卷积核代替 GRU 中的全连接层，即将全连接变为局部连接。ConvGRU 将矩阵运算和卷积操作相结合，能够同时获取时间信息和空间信息。

ConvGRU 的具体定义如式（10）—（13）所示：

$$Z_t = \sigma(W_{xz} * x_t + W_{hz} * H_{t-1}), \tag{10}$$

$$R_t = \sigma(W_{xr} * x_t + W_{hr} * H_{t-1}), \tag{11}$$

$$\widetilde{H}_t = \tanh[W_{xh} * x_t + R_t \circ (W_{hh} * H_{t-1})], \tag{12}$$

$$H_t = (1 - Z_t) \circ H'_t + Z_t \circ H_{t-1}, \tag{13}$$

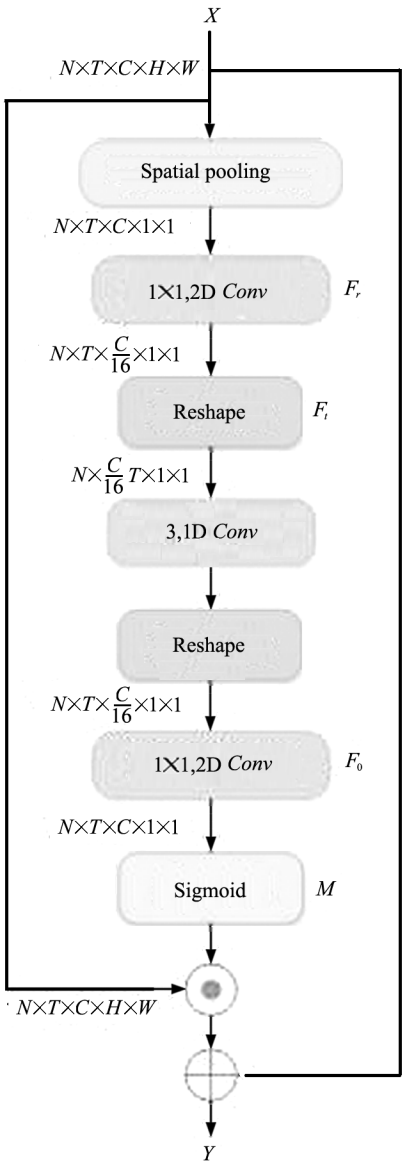

图 3　通道激励模块

Fig. 3　Channel excitation module

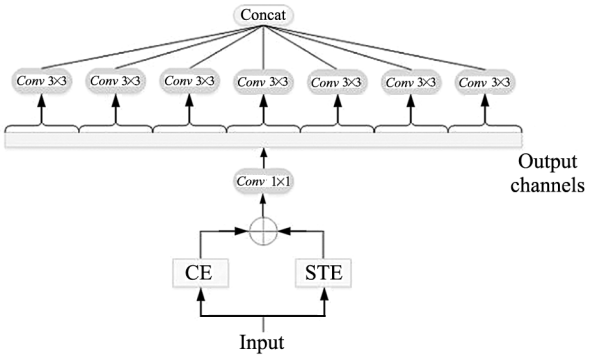

图 4　Xception 最后一个残差块的起始点嵌入 STE 和 CE

Fig. 4　The starting point of the last residual block of Xception is embedded in STE and CE

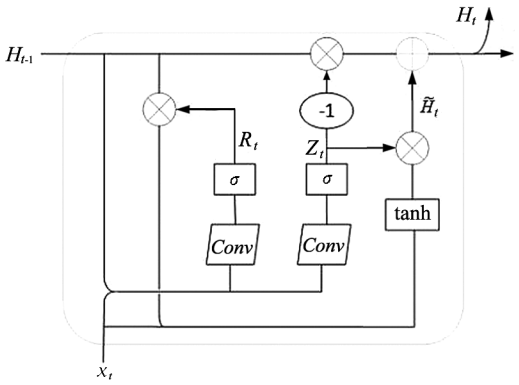

图 5　ConvGRU 网络结构图

Fig. 5　Diagram of ConvGRU network structure

式中，tanh(•) 为双曲正切函数，* 表示卷积操作，。表示元素相乘，Z_t、R_t 分别表示更新门和重置门，H_{t-1}、H_t 和 $\widetilde{H}_t$ 分别表示上一时刻状态、当前状态和候选状态。

1.4　改进的输出分类器

传统的动作识别分类器大多是由多层感知机(multilayer perceptron，MLP)操作进行动作分类的。MLP 由输入层、隐藏层、输出层构成，且层与层之间的连接方式为全连接。该方法仅提供了特征映射的固定线性聚合，忽略了较小目标的特征，对较小目标进行特征提取及检测时，待检测目标容易与其他目标的平均特征混淆从而导致不可识别。因此本文对输出分类器进行改进，将两层全连接层运用 AFF[14] 机制进行连接，以此更好地融合尺度不一致的特征。如图 6 所示为改进前后的输出分类器对比。

在改进的输出分类器中，AFF 模块使用多尺度通道注意模块(multi-scale channel attention mod-

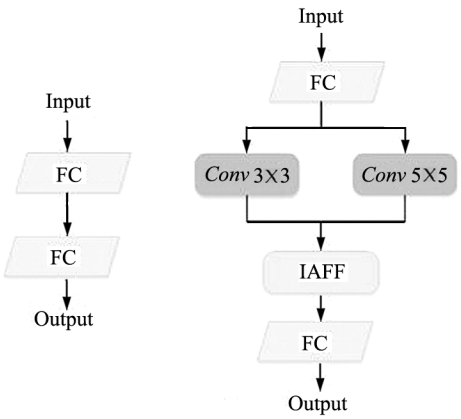

图 6　改进前后的输出分类器对比

Fig. 6　Comparison of output classifiers before and after improvement

ule,MS-CAM)作为其核心。MS-CAM 提出了通道中的尺度问题,并通过点态卷积来实现。

本文在 MS-CAM 模块的基础上,添加不同大小的卷积核,构造了改进的多尺度通道注意模块(MCAM),如图 7 所示。物体和场景的尺度会随着时间的推移而产生较大变化,在此基础上,应用多个不同尺寸的卷积核,相比 MS-CAM,可以提高对不同尺度特征的适应能力,保留更多的局部细节。在 MCAM 模块中,对输入特征图进行通道维数压缩,以减少计算负担。其次,并行执行具有不同尺寸的卷积核运算,其中包含批量归一化和 ReLU 函数处理,将获得的多尺度特征图与压缩输入链接在一起。获取拼接后的特征图有助于探索多尺度信息,并有效地从相邻尺度的特征图中挖掘上下文信息,传输到 MS-CAM 模块中进一步获得多尺度特征。

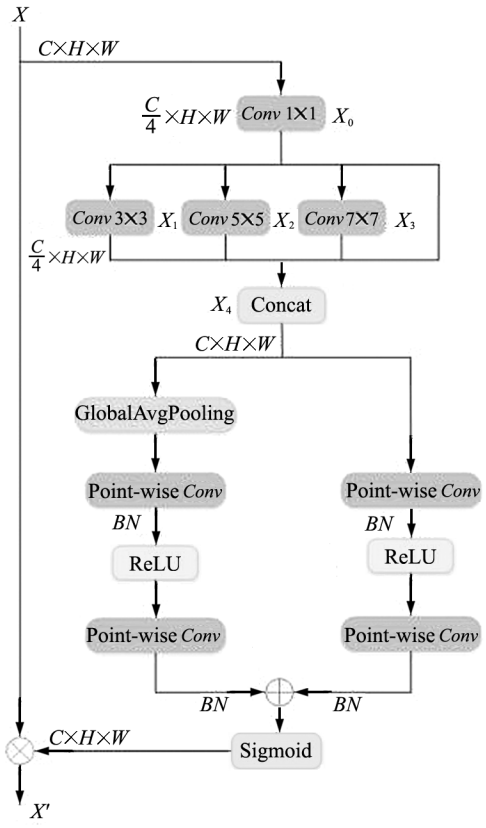

图 7　MCAM 模块

Fig. 7　MCAM module

式(14)—(20)为 MCAM 模块的计算式:

$$X_0 = Conv_{1\times1}(X), \tag{14}$$

$$X_i = Conv_{k_i\times k_i}(X_0), i = 1,2,3, \tag{15}$$

$$X_4 = Cat([X_0, X_1, X_2, X_3]), \tag{16}$$

$$L(X) = B\{PWConv_2\{\delta\{B[PWConv(X_4)]\}\}\}, \tag{17}$$

$$g(X) = GAP[L(X)], \tag{18}$$

$$X' = X \otimes [L(X) \oplus g(X)], \tag{19}$$

式中,$k_i = 2\times i + 1$,$Cat(\cdot)$ 表示 Concat 函数,B 表示 Batch Normalization,$PWConv(\cdot)$ 表示逐点卷积,$\delta(\cdot)$ 表示 ReLU 激活函数,$GAP(\cdot)$ 表示全局平均池化。

本文从特征融合角度出发,提出了一种基于改进的多尺度通道注意力特征融合模块 MCAM-AFF,如图 8 所示。

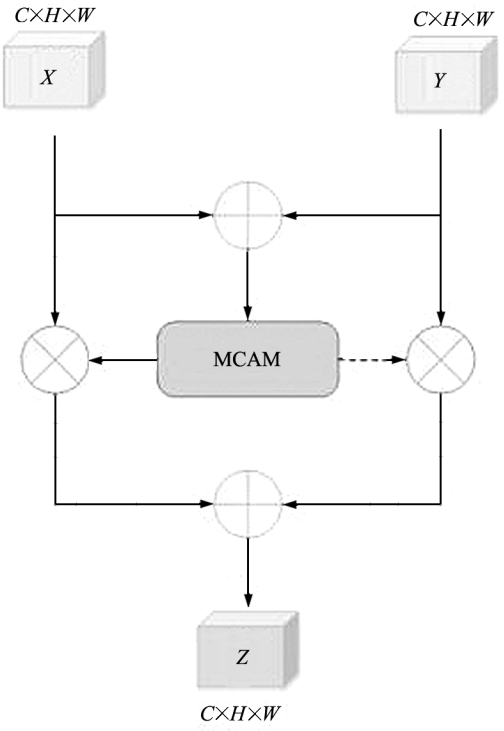

图 8　MCAM-AFF 模块

Fig. 8　MCAM-AFF module

基于多尺度通道注意模块的 AFF 模块计算式如下所示:

$$Z = M(X \oplus Y)\otimes X + [1 - M(X \oplus Y)]\otimes Y, \tag{20}$$

式中,$M(\cdot)$ 表示 MCAM 模块的运算和操作。

2　实验结果与分析

2.1　实验数据集

本文在两个广泛使用的动作识别数据集上评估了模型的性能,即 UCF101 数据集和 HMDB51 数据集。

UCF101 数据集[15]主要来源于 YouTube 视频,有 101 个动作类别,共计 13 320 段视频。该数据集可划分为 5 种类型,分别为人与物体交互的身体动

作、肢体动作、人与人之间交互的身体动作、演奏乐器和运动。

HMDB51 数据集[16]由布朗大学(Brown University)发布,视频主要来源于电影,还有一部分来自公共数据库以及 YouTube 和 Google 等网络视频库。该数据集有 51 个动作类别,共计 6 849 段视频,每个类别中至少包含 101 段视频。该数据集可划分为 5 种类型,分别为面部动作、通过物体操作进行的面部动作、全身动作、与物体交互的身体动作和与人之间交互的身体动作。

2.2　实验设置

实验的硬件环境为 Linux Ubuntu 18.04 系统,显卡为 NVIDIA GeForce GTX 3090。软件环境采用 Pytorch 深度学习框架进行神经网络的训练与测试,batch size 设置为 32,为防止模型训练过拟合,将每个全连接层的的 dropout 设置为 0.5,初始学习率设置为 0.000 1,训练采用 Adam 作为优化器,epoch 次数设定为 50。

本文选取动作识别的准确率作为评估模型识别能力的重要指标。对于视频数据集,将其随机划分为训练集和测试集,比例为 3∶1;并对每个视频段均匀地采样 40 帧,用融入注意力机制的 Xception 对视频帧进行空间特征提取,作为该视频段的特征,再用 ConvGRU 对提取到的序列特征进行时间特征提取,最后通过改进的输出分类器得到识别结果。

2.3　实验结果

2.3.1　定量分析

为验证不同的 CNN 提取空间特征对模型分类效果的影响,采用在 Imagenet 上预训练好的 Inception V3、Inception_Resnet_V2 和 Xception 网络进行对比。将 3 种网络提取到的空间特征分别输入到 ConvGRU 网络获得时间特征,最后输入 Softmax 层获得识别结果。对于不同基础网络的选取,仅在 UCF101 数据集上进行对比分析。

由表 1 可知,Xception 网络效果最佳,识别准确率达到 92.76%,反映出深度可分离卷积运用多个不同尺寸的卷积核,提高对不同尺度特征的适应能力,运用逐点卷积可以在降维或升维的同时,提高网络的表达能力,从而提升了模型的识别准确率,因此在后续实验中选择 Xception 网络作为提取视频空间特征的基础网络。

表 1　不同 CNN 对模型效果的影响

Tab. 1　The influence of different CNNs on model effect

Model structure	Accuracy/%
Inception V3	85.74
Inception_Resnet_V2	90.26
Xception	92.76

神经元参数直接影响网络模型的精度,本文比较了 ConvGRU 网络中 filters 分别为 20、30、40 和 50 时对模型效果的影响。由表 2 可知,在 UCF101 数据集上,filters 为 40 时效果最佳。

表 2　不同 filters 对模型效果的影响

Tab. 2　The influence of different filters on model effect

Filters	Accuracy/%
20	94.71
30	95.13
40	95.66
50	95.58

图 9、图 10 为本文方法在 UCF101 数据集上训练时 *Loss* 值、Top1 准确率和 Top5 准确率随着迭代次数的变化过程。由图可知,大概在迭代次数为 40 次的时候,模型开始收敛,*Loss* 值、Top1 准确率和 Top5 准确率分别收敛至 0.13、95.66% 和 98.21%。

混淆矩阵可以用来可视化地展示分类模型的性能,因此采用混淆矩阵作为评估本文算法的一个重要指标,如图 11 所示。其中,混淆矩阵的横轴和纵轴分别表示 UCF101 和 HMDB51 数据集中各个动作类别的预测标签和真实标签。矩阵对角线上方格的颜色表示算法对该动作类别的识别准确率,对角线以外的方格颜色表示误识别为其他动作的概率。因此,对角线上的方格颜色越深表示算法对该动作类别的识别性能越强。

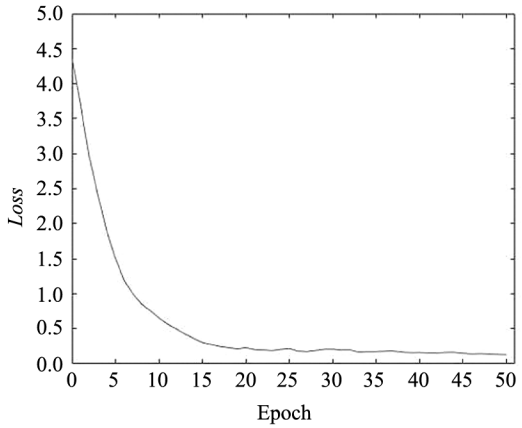

图 9　*Loss* 值随着迭代次数的变化

Fig. 9　*Loss* value changed with the number of epochs

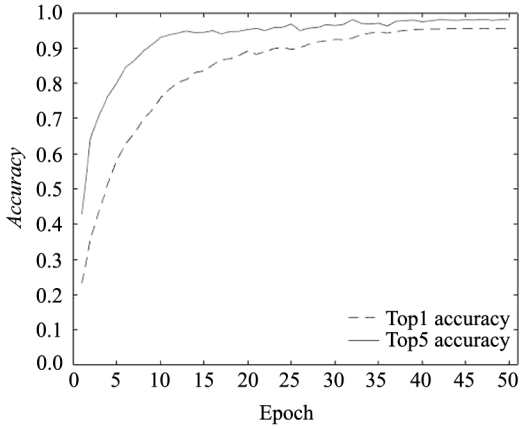

图 10　Top1 准确率和 Top5 准确率随着迭代次数的变化

Fig. 10　Top1 accuracy and Top5 accuracy changed with the number of epochs

由图 11 可知，本文方法在 UCF101 数据集上的识别效果优于 HMDB51 数据集。对于环境和运动轨迹相似的动作，误识别率相对较高，如平衡木和自由体操、笑和微笑等。其他绝大多数的动作类别都能准确识别，误识别率低。

2.3.2　模型整体性能比较

为了体现本文方法的有效性，在 UCF101 数据集和 HMDB51 数据集上，将本文所提出的方法与当前动作识别领域的主流方法进行对比。由表 3 可知，本文方法在 UCF101 数据集和 HMDB51 数据集上取得了不错的结果，分别达到了 95.66 % 和 69.82%的识别准确率，并优于大部分当前的主流方法。表明了在人体动作识别中，空间特征提取部分增强时序建模能力，时间特征提取部分增强空间信息提取能力，二者互为补充，得到更加完整的时空特征，并改进的输出分类器对于模型整体性能有较好的提升效果。

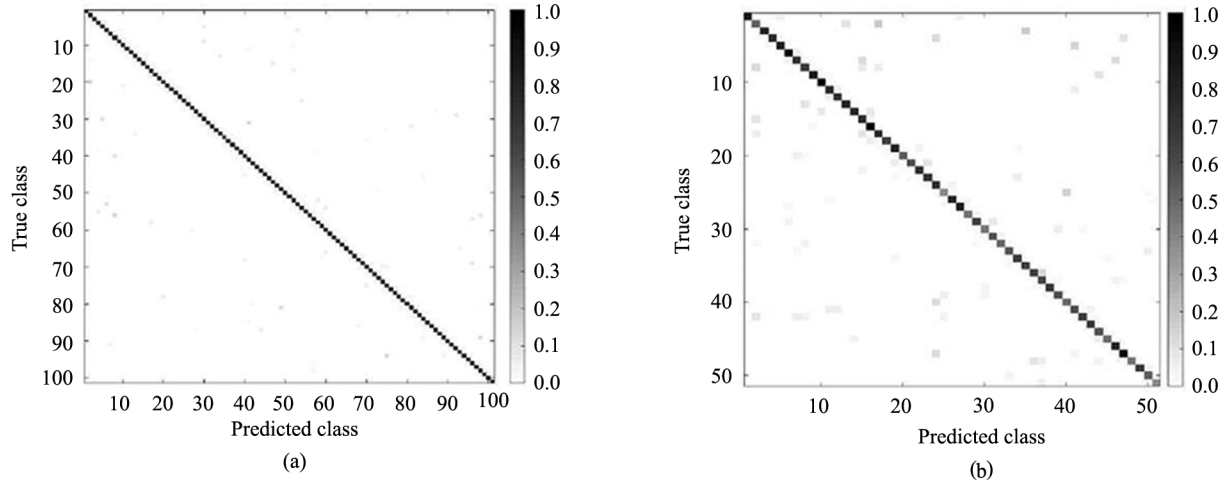

图 11　基于 UCF101 和 HMDB51 数据集的混淆矩阵：(a) UCF101 数据集；(b) HMDB51 数据集

Fig. 11　Confusion matrix based on UCF101 and HMDB51 datasets: (a) UCF101 dataset; (b) HMDB51 dataset

表 3　不同模型在 UCF101 和 HMDB51 数据集上的准确率比较

Tab. 3　Comparison of accuracy of different models on UCF101 and HMDB51 datasets

Model	UCF101/%	HMDB51/%
LRCN[7]	87.9	—
Two-stream[5]	88.0	59.4
STC-ResNet101[17]	93.7	66.8
STS[18]	93.9	67.2
STAM[19]	94.3	69.1
Two-stream CNN[20]	88.0	59.4
Ours	95.66	69.82

2.3.3　消融实验

为进一步验证每个模块对模型整体性能的影响及模型的有效性，选择在 UCF101 数据集上开展消融实验进行对比分析，从表 4 可以看出，将“CNN＋LSTM”模型结构替换为“CNN＋ConvGRU”后，准确率提升了 4.29%，验证了 ConvGRU 结构优于 LSTM 结构；将时空域和通道域的注意力机制融入 Xception 网络中后，识别准确率提升了 0.82%，验证了在空间特征提取阶段获取时序特征的重要性；将 AFF 融入模型输出层后，准确率提升了 1.13%，表明了特征融合在输出分类器中可以得到细小的特征；将 AFF 模块进行改进后得到的 MCAM-AFF 模块使准确率提升了 0.95%。由此可以得出结论，引入了 ConvGRU 模块、SC-Xception 模块和 MCAM-AFF 模块对提升整体模型的准确率是有效的。

表 4　不同模块对整体性能的影响

Tab. 4　The impact of different modules on the overall performance

Model structure	*Accuracy*/%
Xception+LSTM	88.47
Xception+ConvGRU	92.76
SC-Xception+ConvGRU	93.58
SC-Xception + ConvGRU + AFF	94.71
SC-Xception + ConvGRU + MCAM-AFF	95.66

3　结　论

为获得更加完整的时空特征,本文提出了一种融入注意力机制的 Xception 模块、ConvGRU 模块和 MCAM-AFF 模块相结合的模型。该方法通过使用融入注意力机制的 Xception 模块提取视频帧的空间局部信息,ConvGRU 模块提取时序上的全局信息,以获取人体动作的时空特征,最后将提取到的时空特征输入到改进的输出分类器获得识别结果。实验结果表明,该模型在 UCF101 和 HMDB51 数据集上分别达到了 95.66%和 69.82%的分类准确率,得到了较好的结果。在目前主流的动作识别模型中,对于视频背景和轨迹相似的不同类别的动作,有一定的误识别率,本文方法降低了一定的误识别率,但依然存在此类问题。因此,在未来的工作中,将进一步考虑对网络的改进,提升识别准确率,降低误识别率,优化模型的整体性能。

参考文献:

[1] AHMAD T, JIN L W, ZHANG X, et al. Graph convolutional neural network for human action recognition: a comprehensive survey[J]. IEEE Transactions on Artificial Intelligence, 2021, 2(2): 128-145.

[2] HU J L, QI Y F, WANG J Y. Student behavior recognition in online classroom based on spatiotemporal graph convolutional network[J]. Journal of Optoelectronics · Laser, 2022, 33(2): 149-156.
胡锦林,齐永锋,王佳颖. 基于时空图卷积网络的学生在线课堂行为识别[J]. 光电子 · 激光, 2022, 33(2): 149-156.

[3] WANG H, SCHMID C. Action recognition with improved trajectories[C]//IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE, 2013: 3551-3558.

[4] QIAN H F, YI J P, FU Y H. Review of human action recognition based on deep learning[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(3): 438-455.
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.

[5] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[EB/OL]. (2014-06-09)[2023-03-20]. https://arxiv.org/abs/1406.2199.

[6] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 4489-4497.

[7] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 2625-2634.

[8] PAN B, SUN J, LIN W, et al. Cross-stream selective networks for action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, June 16-17, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 454-460.

[9] JAOUEDI N, BOUJNAH N, BOUHLEL M S. A new hybrid deep learning model for human action recognition[J]. Journal of King Saud University-Computer and Information Sciences, 2020, 32(4): 447-453.

[10] ZHAO D G, ZHI M. Multi-scale spatiotemporal graph convolution algorithm for human motion recognition[J]. Journal of Computer Science and Exploration, 2023, 17(3): 719-732.
赵登阁,智敏. 用于人体动作识别的多尺度时空图卷积算法[J]. 计算机科学与探索, 2023, 17(3): 719-732.

[11] WANG Z, SHE Q, SMOLIC A. Action-net: multipath excitation for action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, USA. New York: IEEE, 2021: 13214-13223.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.

[13] SHI X J, GAO Z H, LAUSEN S L, et al. Deep learning for precipitation nowcasting: a benchmark and a new model [C]//Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA. Red Hook, NY, USA: Curran Associates Inc., 2017: 5617-5627.

[14] DEL RIO R E, PARDO-NOVOA J C, CERDAGARCIA-ROJASC M, et al. Vibrational circular dichroism behavior of quinol cacalolides from Psacalium aff. sinuatum[J]. Journal of Molecular Structure, 2021, 1224: 128987.

[15] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human action classes from videos in the wild[EB/OL]. (2012-12-03) [2023-03-21]. https://arxiv.org/abs/1212.0402.

[16] KUEHNE H, JHUANG H, GARROTE H, , et al. HMDB: A large video database for human motion recognition[C]// IEEE International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE, 2011: 2556-2563.

[17] DIBA A, FAYYAZ M, SHARMA V, et al. Spatio-temporal channel correlation networks for action classification [C]//European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Cham: Springer, 2018: 299-315.

[18] TU Z, XIE W, DAUWELS J, et al. Semantic cues enhanced multimodality multistream CNN for action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(5): 1423-1437.

[19] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding[C]//IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE, 2019: 7082-7092.

[20] DHIMAN C, VISHWAKARMA D K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics[J]. IEEE Transactions on Image Processing, 2020, 29: 3835-3844.

作者简介：

张荣芬 (1977—)，女，博士，教授，硕士生导师，主要从事机器视觉、智能算法和智能硬件等方面的研究.