2025 年 第 25 卷 第 4 期
2025, 25(4): 01540-07

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2401484
引用格式:曹若琛, 冯秀芳, 赵晨. 基于运动特征增强双流网络的视频行为识别[J]. 科学技术与工程, 2025, 25(4): 1540-1546.
Cao Ruochen, Feng Xiufang, Zhao Chen. Video action recognition based on sport feature enhancement in two-stream network[J]. Science Technology and Engineering, 2025, 25(4): 1540-1546.

基于运动特征增强双流网络的视频行为识别

曹若琛, 冯秀芳, 赵晨*
(太原理工大学软件学院, 晋中 030600)

摘 要 为解决目前行为识别中双流网络对运动特征提取的不充分导致识别准确度低的问题,通过提出基于运动特征增强双流网络的行为识别方法提高准确率。该网络分为空间流和时间流,空间流网络和时间流网络结构相同,输入不同。空间流网络输入为视频帧序列,而时间流网络输入为视频帧差序列。网络结构以 Resnet50 为骨干网络,将 3×3 卷积替换为所提出的全局运动特征模块和局部运动特征模块,充分提取视频运动信息,最终将空间流和时间流结合输出结果。结果表明:该模型在 UCF101 和 HMDB51 数据集上准确率达到 96.8% 和 75.3%,与传统算法相比有一定优越性。

关键词 行为识别; 深度学习; 运动特征; 双流网络
中图法分类号 TP391; **文献标志码** A

Video Action Recognition Based on Sport Feature Enhancement in Two-stream Network

CAO Ruo-chen, FENG Xiu-fang, ZHAO Chen*
(School of Software, Taiyuan University of Technology, Jinzhong 030600, China)

[**Abstract**] To solve the problem of insufficient extraction of sport features by dual stream networks in current action recognition, which leads to low recognition accuracy, a action recognition method based on sport feature enhancement two-stream networks was proposed to improve accuracy. The network was divided into spatial stream and temporal stream, with the same structure but different inputs. The input of the spatial stream network was a video frame sequence, while the input of the temporal stream network was a video frame difference sequence. The network structure used Resnet50 as the backbone network, replacing the 3×3 convolution with the proposed global sport feature module and local sport feature module, fully extracting video sport information, and finally combining spatial and temporal stream to output the results. The results show that the accuracy of the model on the UCF101 and HMDB51 datasets reaches 96.8% and 75.3%, which is superior to traditional algorithms.

[**Keywords**] action recognition; deep learning; sport feature; two-stream network

随着人工智能迅速发展,计算机视觉取得了一系列研究成果,具有时间维度的视频行为识别领域得到广泛关注[1-3]。当前,视频行为识别方法分为传统方法和深度学习方法。相比传统方法,深度学习方法准确率更高且泛化能力更强,但仍存在缺陷。深度学习方法主要有双流卷积网络[4-6],3D 卷积网络[7-9]以及 transformer 网络[10-13]。双流卷积网络相比 3D 卷积网络和 transformer 网络复杂度更低,不易过拟合,因此主要围绕双流卷积进行。

双流卷积网络由 Simonyan 等[14]于 2014 年提出,该方法解决了早期二维卷积无法提取时间信息的问题,提出对空间信息和时间信息分别提取再进行融合的方法来提取视频特征。双流网络分为空间流网络和时间流网络,其中,空间流网络输入图像用来提取视频空间信息,时间流网络通过输入光流提取时间信息,最后对两个网络的结果进行融合。虽然这种早期双流网络通过融合时间流获取视频时间信息,但其不能进行长时序建模,准确率较低,因此目前视频行为识别主要应用 TSN(temporal segment networks)[15]和 TSM(temporal shift module)[5]这种支持长时序建模的双流网络。

TSN 和 TSM 是经典的双流卷积模型。其中,TSN 对长时序连续视频帧进行等分,接着对每个等分片段提取特征,最终对多个结果进行加权平均。

收稿日期: 2024-03-04; **修订日期**: 2024-11-20
基金项目: 山西省重点研发计划(202102020101007)
第一作者: 曹若琛(1990—),男,汉族,山西太原人,博士,讲师。研究方向:虚拟现实、行为识别。E-mail:caory004@163.com。
* **通信作者**: 赵晨(1998—),男,汉族,辽宁沈阳人,硕士研究生。研究方向:行为识别。E-mail:906070493@qq.com。

TSM 则是融合了时间位移模块,通过在时间维度移动像素获取时间特征。TSN 和 TSM 都采用长视频序列中截取若干帧的方式解决了传统双流网络无法对长时序进行建模的问题。但目前的双流网络仍存在对运动特征信息提取不充分的问题。TSN 网络分段采样策略提取特征,但网络并未高效提取运动特征。TSM 在 TSN 基础上加入时间位移模块高效提取了局部运动特征,但缺乏全局运动特征建模。丁雪琴等[4]提出时空异构双流卷积网络去除冗余特征信息,但该方法仍未能高效提取运动特征。鉴于此,提出一种运动特征增强双流网络(sport feature enhancement two-stream network,SFETN),通过引入本文模块取得较高行为识别精度;设计一种全局运动特征模块(global sport feature module,GSFM),通过扩大特征图差异,高效提取视频全局运动特征;设计一种局部运动特征模块(local sport feature module,LSFM),通过改进时间位移操作,高效提取视频局部运动特征,在双流网络中实现对运动特征的充分提取以提高行为识别准确率,以期通过增强并融合视频全局和局部运动特征,解决双流网络模型特征提取不充分的问题。

1　运动特征增强双流网络

1.1　整体架构

基于双流卷积结构,提出运动特征增强双流网络,整体框架如图 1 所示。

为了能够实现对视频的长时序建模,本文网络借鉴 TSN 网络思想,对长视频序列进行间隔片段抽取,同时时间流输入使用帧差(视频帧相减)代替光流,避免数据预处理时光流抽取带来的高计算和长耗时。网络首先将一段长序列视频帧分为若干个视频片段,然后在每个视频片段中选择一帧 RGB 视频图像和相邻视频帧的帧差图像,RGB 图像通过空间流网络进行空间特征提取,帧差图送入时间流网络进行时间特征提取,依次对每个视频片段都进行如上操作,这样就得到了若干个空间流网络分类结果和若干个时间流网络分类结果。对这些空间流网络分类结果做 Segmental Consensus (一种共识)得到空间流最终结果,对相应的时间流网络分类结果做 Segmental Consensus 得到时间流最终结果,最后将空间流分类结果和时间流分类结果作加权融合得到最终精度。本文网络建模方法如式(1)、式(2)所示。

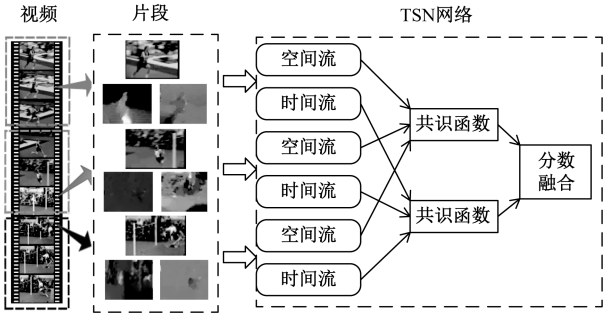


图 1　网络整体框架

Fig. 1　Overall framework of the network

$$G = g[F(T_1;w),F(T_2;w),\cdots,F(T_k;w)] \tag{1}$$

$$L(y,G) = -\sum_{i=1}^{c} y_i\left(G_i - \lg\sum_{j=1}^{c}\exp G_j\right) \tag{2}$$

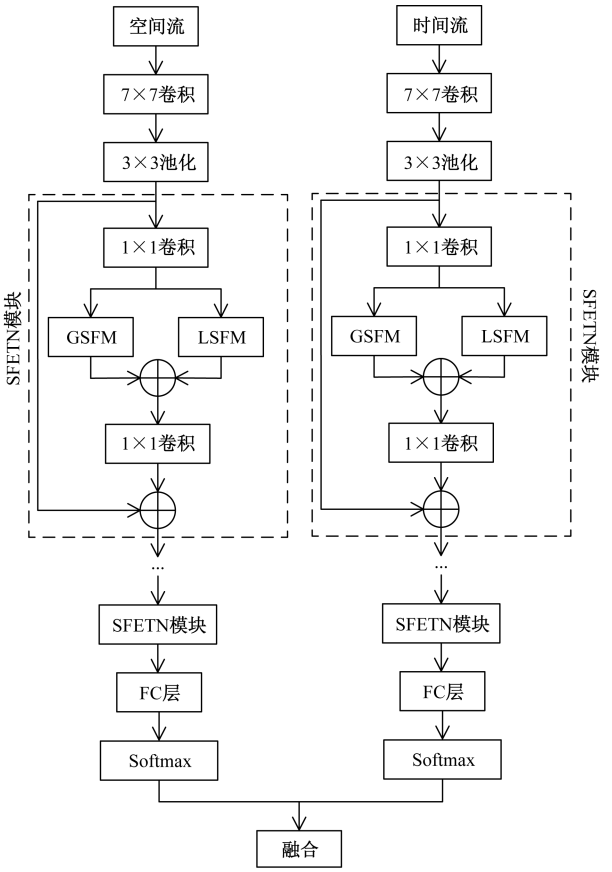
式中:$(T_1,T_2,\cdots,T_k)$ 为视频片段;$F(T_k;w)$ 为参数为 w 的网络;g 为融合函数,对同一类得分取平均值;G 为段共识函数,结合交叉熵损失得到最终损失函数 L;c 为类别数;y 为每类的值;y_i 为类别 i 的真实标签;G_i 为类别 i 的分数;G_j 为类别 j 的分数。

1.2　网络结构

本文网络分为空间流网络和时间流网络,空间流网络和时间流网络结构相同,输入不同。空间流输入为视频帧序列,时间流输入为帧差序列。网络结构使用 Resnet50 为骨干网络,提出全局运动特征模块 GSFM 和局部运动特征模块 LSFM 替换掉原 Resnet50 网络模块中的 3×3 卷积,以更加充分地提取视频的运动特征,网络结构如图 2 所示。

空间流网络输入为视频帧序列,最初经由一个 7×7 的卷积和 3×3 的最大池化来进行特征提取并减少计算量去除冗余信息。随后通过若干层 SFETN 模块进行特征提取。由于本文网络使用 Resnet50 为骨干网络,因此 SFETN 模块的数量对应 Resnet50 网络中卷积模块的数量。每个 SFETN 模块的结构在图 2 中所示,特征图进入 SFETN 模块时,首先会经过 1×1 卷积减小维度,缩小计算量。然后分为两个分支,一个分支经过全局运动特征模块 GSFM 提取视频的全局运动特征,使得网络学习到视频的整体关联性;另一个分支经过局部运动特征模块 LSFM 提取视频局部运动特征,使得网络能够关注到视频的局部关键运动。将两个分支相加后,网络整体能够充分学习到视频的运动信息,从而提高行为识别的准确率。在经过若干 SFETN 模块提取特征后,最后经由 FC 层和 Softmax 函数得到空间流的分类结果。时间流输入为视频帧差序列,经过同样的网络结构提取相应的时间信息后,经由 FC 层和 Softmax 函数得到时间流的分类结果。最终将空间流分类结果和时间流分类结果加权融合得到最终结果,如式(3)所示。

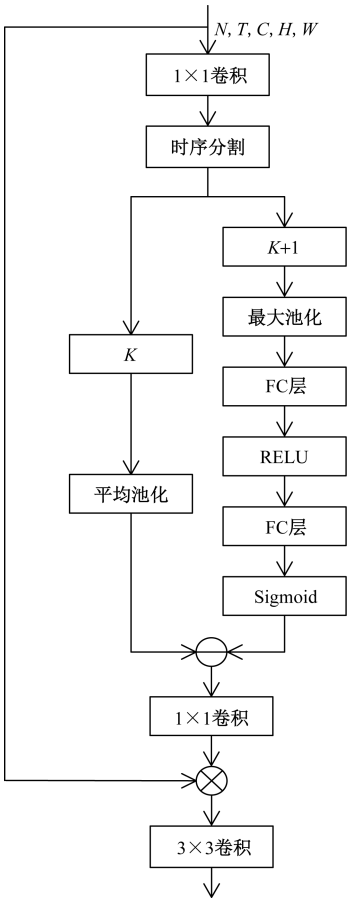
$$R = \sum_{i=1}^{n}\frac{xA_i + y_{\text{time}}B_i}{n} \tag{3}$$



FC 为全连接层

图 2 网络结构

Fig. 2 Network structure



K 为特征图序号;RELU 和 Sigmoid 为激活函数

图 3 全局运动特征模块结构

Fig. 3 Global sport feature module structure

式(3)中: A_i 为空间流每一类分数; B_i 为时间流每一类分数; x 为空间流权重; y_{time} 为时间流权重; n 为类别数; R 为分类结果。

1.2.1 全局运动特征模块

提出全局运动特征模块 GSFM 提取视频的全局运动特征,以解决目前网络对全局运动特征提取不充分的问题,其结构如图 3 所示。

全局运动特征可以通过特征图之间的整体差异进行实现。模块输入特征为 X,特征尺寸为 $[N,T,C,H,W]$,其中, N 为批处理视频帧数, T 为时间维度数, C 为通道数, H 和 W 分别为特征图的长度和宽度。特征图首先经过 1×1 卷积进行通道下采样,目的是减少后续计算量,可表示为

$$X^m = \text{Conv}(X),\quad X^m \in \mathbf{R}^{N\times T\times C_1\times H\times W} \tag{4}$$

式(4)中:Conv 为 1×1 卷积; X^m 为通道下采样后的特征; C_1 为通道下采样后的通道数。

在通过通道下采样后,特征图会在时间维度上进行等时序分割,这里以 K 和 $K+1$ 示意时序分割后的两个特征图,进行特征帧作差处理。但是由于每个特征帧之间差异不够明显,为了增大相邻特征帧之间的差异,对相邻帧的后一帧进行通道增强后减去前一帧进行处理。具体为对于特征尺寸为 $C_1\times H\times W$ 的第 $K+1$ 帧特征,首先使用最大池化,使用最大池化的目的是放大不同帧之间的全局运动特征差异,同时压缩空间特征,将特征尺寸变为 $C_1\times 1\times 1$,使得网络获得一个全局描述,可表示为

$$z = \frac{1}{HW} = \sum_{i=1}^{H}\sum_{j=1}^{W} Y_{ijc},\quad c \in [1,C_1] \tag{5}$$

式(5)中: Y 为 $K+1$ 帧特征图; z 为输出结果。

特征图经过最大池化特征尺寸变为 $C_1\times 1\times 1$ 后,由于长和宽被压缩为 1,会使网络更关注到通道级别的特征。之后使用一个 FC 层降维,以降低计算量并增强关键特征。之后使用 RELU 函数增强网络各层之间的非线性关系,提高模型表达能力。再使用 FC 层升维,还原特征信息,最后通过 sigmoid 函数特征归一化得到特征图 D_1。对于特征尺寸为 $C_1\times H\times W$ 的第 K 帧特征,使用平均池化压缩空间特征,使得特征尺寸变为 $C_1\times 1\times 1$,得到特征图 D_2,这里使用平均池化而非最大池化的目的是让两帧相减后的全局运动特征差异达到最大,从而让网络能够更加明显地关注到运动差异信息。在得到

D_1 和 D_2 后,将两特征图进行相减,得到全局运动特征 D_0,可表示为

$$D_0 = D_1 - D_2 \tag{6}$$

将所有时序分割后的特征图都进行如上操作,然后将得到的帧差特征图进行拼接,此时特征时间维度为 $T-1$,为了保持时间维度和输入时相同,用零表示时间步结束的运动特征,再经过 1×1 卷积还原通道数为 C,得到大小为 $[N,T,C,1,1]$ 的特征 D。为还原原特征图大小,将特征 D 与和最初的特征 X 进行相乘,得到最终的原特征图大小 $[N,T,C,H,W]$ 的全局运动特征,最后使用 3×3 卷积对空间特征进行增强,进一步提升网络性能。

1.2.2　局部运动特征模块

提出能够高效对局部运动特征建模的局部运动特征模块 LSFM,使用位移算子重排列思想,沿时间维度移动像素,使得视频帧拥有时间信息,同时使用卷积增强空间特征,提升性能。

局部运动特征模块由改进时间位移模块和 3×3 卷积构成,如图 4 所示。改进时间位移模块借鉴了时间位移模块 TSM,并在其基础上进行改进。传统网络在时间维度上独立进行特征提取,因此没有时间建模效果(图 5 左侧)。时间位移模块简单来说就是令当前帧混合其相邻帧的像素信息,使其能够感知到时间信息。以卷积核大小为 3 的卷积过程为例,设输入 $\boldsymbol{X}$ 为一维向量,权值 $\boldsymbol{W}=(w_1,w_2,w_3)$,则进行时间位移的卷积操作可以分为两个过程,首先沿时间维度移动特征,可表示为

$$X_i^{-1} = X_{i-1},\ X_i^0 = X_i,\ X_i^{+1} = X_{i+1} \tag{7}$$

进行乘法累加,乘以权重 $\boldsymbol{W}$ 得

$$\boldsymbol{Y} = w_1\boldsymbol{X}^{-1} + w_2\boldsymbol{X}^0 + w_3\boldsymbol{X}^{+1} \tag{8}$$

上述过程为一维向量时间位移卷积过程。可以看到,在没有增加任何参数量和计算量的情况下,就使得特征图拥有了局部运动特征。图 5 左侧表示含有通道维度和时间维度的张量,其中每一行代表不同时间特征,之后进行时间位移后,如图 5 右侧所示,将部分通道沿时间维度进行移动,这样对

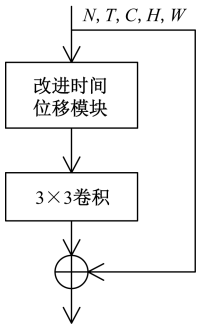


图 4　局部运动特征模块结构

Fig. 4　Local sport feature module structure

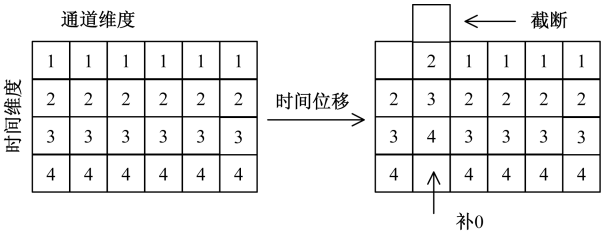


图 5　时间位移模块

Fig. 5　Temporal shift module

于每一行时间特征,就拥有其他通道的时间信息,即有表示一定运动特征的能力,但由于仅移动少部分通道,因此只能表示局部运动特征。

本文改进时间位移模块受到 TSM 的启发,使用变化通道代替固定的移位通道,假设输入维度为 $[N,T,C,H,W]$,对于 TSM 移位过程,移位通道数固定为 $C/4$,其他通道固定不动,可表示为

$$\begin{cases} [N,t,C,H,W] = [N,t+1,C,H,W], \\ \qquad 1 \leqslant t \leqslant T-1,\ 1 \leqslant c \leqslant \dfrac{C}{8}, \text{下移} \\ [N,t,C,H,W] = [N,t-1,C,H,W], \\ \qquad 2 \leqslant t \leqslant T,\ \dfrac{C}{8} \leqslant c \leqslant \dfrac{C}{4}, \text{上移} \\ [N,t,C,H,W] = [N,t,C,H,W], \\ \qquad 1 \leqslant t \leqslant T,\ \dfrac{C}{4} \leqslant c \leqslant C, \text{不动} \end{cases} \tag{9}$$

由于 TSM 移动通道固定,因此对于局部特征的关注过于单一,缺乏特征多样性,因此为了改进这个问题,所提出的改进时间位移模块将固定通道改进为变化的通道,将移动通道数随机设为 $\frac{C}{4}$ 或 $\frac{C}{2}$,并将移动通道的选取也进行随机化,通过以上操作,能够让网络提取的局部特征信息更加全面,再经过时间移位后,通过 3×3 卷积提取空间特征进一步提升网络性能,最后使用残差连接方式,将此时特征与模块输入时的特征进行相加得到最终结果。

2　实验

2.1　数据集

在 UCF101[16] 和 HMDB51[17] 上进行消融和对比实验。鉴于 HMDB51 数据集规模较小且质量有限,将 UCF101 数据集作为主要数据集进行消融实验,而将 HMDB51 作为辅助数据集仅用在对比实验中。UCF101 数据集中视频取自 YouTube 网站,共包含 13 320 个视频和 101 个类别。这些类别包含人与物交互,人和人互动,体育,乐器演奏等。每种动作分为 25 组,每组包含 4~7 个含有相似特征的视频。所有视频共 6.5 GB,分辨率为 320×240,帧率为 25 FPS。UCF101 的详细信息如表 1 所示。

表 1 UCF101 数据集信息

Table 1 UCF101dataset information

信息	数值
运动	101
帧数	13 320
组数	25
每组帧数	4 ~ 7
视频时间/min	1 600
帧率/FPS	25
尺寸	320 ×240

HMDB51 数据集包含个人动作,多人动作,人和物交互等 51 个类别,共有 6 849 个视频。视频分辨率为 320 ×240。数据集收集自 YouTube 和谷歌等视频网站。使用官方划分的数据集结构进行实验。使用训练集训练网络并在测试集上验证网络性能。

2.2 实验环境

本实验在 64 位 Ubuntu 系统下实现,计算机配置为 Intel(R) Xeon(R) Platinum 8255C @ CPU 2.00 GHz,Nvidia GeForce RTX 3090 24 G 显卡,基于 Python 3.8 + Pytorch 1.7.0 + CUDA 11.0 + CUDNN 7.6 实现。对空间流网络和时间流网络分开训练,之后对两个结果进行最大值融合。空间流网络使用 Kinetics400 作为预训练数据集,时间流网络使用 ImageNet 作为预训练数据集。长时序视频分割段数设置为 7,batch size 设置为 16,学习率设置为 0.001,每 10 个 epoch 衰减为原来的 1/10。图像尺寸设置为 224 ×224。模型一共训练 30 个 epoch。

2.3 消融实验

为验证本文全局运动特征模块 GSFM 和局部运动特征模块 LSFM 的作用,以及探究在什么情况下模型性能达到最佳,在 UCF101 上进行消融实验。

首先验证全局运动特征模块 GSFM 和局部运动特征模块 LSFM 是否能够高效提取视频的全局运动特征和局部运动特征。模型采用 Resnet50 为骨干网络,因此对比基础骨干网络为 Resnet50,然后分别使用 GSFM 和 LSFM 对 Resnet50 中的 3 ×3 卷积进行替换,具体为全局运动特征模块替换 3 ×3 卷积、局部运动特征模块替换 3 ×3 卷积、以及全局运动特征模块和局部运动特征模块并联后替换 3 ×3 卷积,分别在空间流网络和时间流网络上进行实验,结果如表 2 所示。

表 2 模块消融实验

Table 2 Module ablation experiment

网络结构	空间流	时间流	双流融合
原结构	92.0	91.3	94.6
GSFM	93.5	92.6	95.3
LSFM	94.6	93.8	96.1
GSFM + LSFM	95.3	94.6	96.8

由表 2 可知,无论是空间流还是时间流,所提的全局运动特征模块和局部运动特征模块都能使得模型准确率得到提升,并且在同时加入两个模块后准确率最佳。

为验证本文改进时间位移模块的有效性,对局部运动特征模块结构分别试用原时间位移模块和本文时间位移模块,实验结果如表 3 所示。可以看出,使用所提出的改进时间位移模块更能捕捉局部运动特征,准确率更高。

对全局运动特征模块 GSFM 和局部运动特征模块 LSFM 的连接方式进行实验,目的是探究两个模块如何能够最大程度提升准确率,实验分 3 种情况,分别为全局运动特征模块串联局部运动特征模块(模块串联 1)、局部运动特征模块串联全局运动特征模块(模块串联 2)和全局运动特征模块并联局部运动特征模块(模块并联),实验结果如表 4 所示,可以看出,当使用并联方式连接模块时,网络性能最佳。

为进一步提升网络识别准确率,对空间流和时间流结果融合方式进行实验,分别实验 3 种融合方式,分别为最大值融合,平均融合以及加权融合,结果如表 5 所示,可以看出,网络在加权融合方式下准确率最高。

表 3 时间位移模块实验

Table 3 Temporal shift module experiment

网络结构	空间流	时间流	双流融合
原结构	92.0	91.3	94.6
GSFM + TSM	94.9	94.3	96.6
GSFM + LSFM	95.3	94.6	96.8

表 4 模块连接方式实验

Table 4 Module connection experiment

网络结构	空间流	时间流	双流融合
模块串联 1	94.9	94.0	96.5
模块串联 2	94.7	93.9	96.3
模块并联	95.3	94.6	96.8

表 5 模型融合方式实验

Table 5 Model fusion method experiment

融合方式	准确率/%
最大值融合	96.3
平均融合	96.6
加权融合	96.8

2.4 对比实验

本文模型分别在 UCF101 和 HMDB51 数据集上进行实验,对比视频动作识别领域的其他方法,其中包括传统方法以及深度学习方法,结果如表 6 所示。

由表6可知,本文网络在UCF101和HMDB51数据集上准确率分别达到96.8%和75.3%,优于其他网络。相对于传统方法(如DVSR、IABC等),本文准确率明显更高。相对于双流网络(如TSM、SHT-CNN等),本文模型使用时间流使用帧差代替光流,避免了光流抽取,同时在参数量和浮点数相差不过大的情况下,准确率更高。相对于3D网络(如X3D、Eco等)和Transformer网络(如Timesformer、SVT等),本文网络参数量和浮点数远低于这些网络,同时准确率更高。

为直观展示本文模型的性能优势,从UCF101数据集中抽取部分视频片段在TSN和本文模型上进行可视化推断,结果如图6所示。可以看出,TSN错误地将冲浪运动识别为皮划艇运动,而本文模型全部识别正确,性能更高。

表6　不同方法性能对比

Table 6　Performance comparison of different methods

方法	参数量/10^6	浮点数/GFLOPs	UCF101	HMDB51
DVSR[18]	—	—	82.9	56.0
IABC[19]	—	—	86.2	62.4
SHT-CNN[4]	—	—	92.3	62.5
TSN[15]	21.4	32.0	94.0	68.5
TSM[5]	48.6	66.0	95.9	73.5
TS-D3D[20]	17.5	29.4	96.0	65.6
Timesformer[10]	121.4	590.0	92.1	—
MFI[21]	24.6	67.2	95.6	—
SVT[11]	173.2	590.0	93.7	67.2
X3D[7]	150.2	143.5	90.3	68.2
AR3D[8]	139.5	125.0	94.8	69.3
Eco[22]	150.0	267.0	94.8	72.4
本文方法	54.8	69.0	96.8	75.3

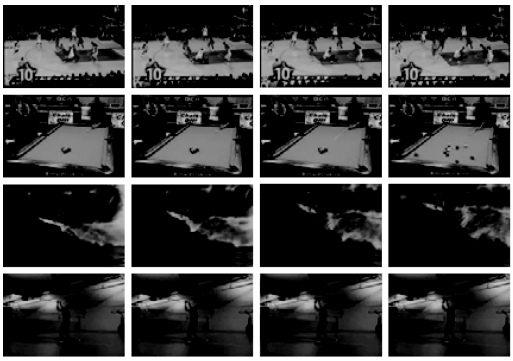


图6　部分可视化推断图

Fig. 6　Partial visual inference graph

3　结论

所提出运动特征增强双流网络行为识别方法。设计全局运动特征模块GSFM和局部运动特征模块LSFM解决双流网络提取运动特征不足的问题并提升识别准确率。同时,模型使用帧差代替光流避免数据预处理阶段额外的计算和耗时。本文算法在视频数据集UCF101和HMDB51上的识别Top-1精度分别达到96.8%和75.3%,在行为识别任务上取得了显著优势。未来的工作将致力于进一步提升行为识别准确率,并尝试探索更轻量化的结构来满足实际应用需求。

参　考　文　献

[1] 袁首, 乔勇军, 苏航, 等. 基于深度学习的行为识别方法综述[J]. 微电子学与计算机, 2022, 39(8): 1-10.
Yuan Shou, Qiao Yongjun, Su Hang, et al. A review of behavior recognition methods based on deep learning[J]. Microelectronics and Computer Science, 2022, 39(8): 1-10.

[2] Kong Y, Fu Y. Human action recognition and prediction: a survey[J]. International Journal of Computer Vision, 2022, 130(5): 1366-1401.

[3] 梁绪, 李文新, 张航宁. 人体行为识别方法研究综述[J]. 计算机应用研究, 2022, 39(3): 651-660.
Liang Xu, Li Wenxin, Zhang Hangning. Summary of research on human behavior recognition methods[J]. Computer Application Research, 2022, 39(3): 651-660.

[4] 丁雪琴, 朱轶昇, 朱浩华, 等. 基于时空异构双流卷积网络的行为识别[J]. 计算机应用与软件, 2022, 39(3): 154-158.
Ding Xueqin, Zhu Yisheng, Zhu Haohua, et al. Behavior recognition based on spatio-temporal heterogeneous double-stream convolution network[J]. Computer Applications and Software, 2022, 39(3): 154-158.

[5] 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. New York: IEEE, 2019: 7083-7093.

[6] 金博, 王景林, 刘泓铄, 等. 基于双注意力时间卷积网络的人体行为识别[J]. 计算机工程与设计, 2023, 44(1): 284-291.
Jin Bo, Wang Jinglin, Liu Hongyue, et al. Human behavior recognition based on dual attention time convolution network[J]. Computer Engineering and Design, 2023, 44(1): 284-291.

[7] Feichtenhofer C. X3D: expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 203-213.

[8] Dong M, Fang Z, Li Y, et al. AR3D: attention residual 3D network for human action recognition[J]. Sensors, 2021, 21(5): 1656.

[9] 闫雨寒, 陈天, 刘忠育, 等. 基于双重注意力和3DResNet-BiLSTM行为识别方法[J]. 计算机应用与软件, 2023, 40(2): 192-196, 205.
Yan yuhan, Chen Tian, Liu Zhongyu, et al. Behavior recognition

method based on dual attention and 3DResNet-BiLSTM[J]. Computer Applications and Software, 2023, 40(2): 192-196, 205.

[10] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? [C]//International Conference of Machine Learning. New York: IEEE, 2021: DOI: 10.48550/arXiv.2102.05095.

[11] Ranasinghe K, Naseer M, Khan S, et al. Self-supervised video transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 2874-2884.

[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 York: IEEE, 2022: 3202-3211.

[13] 姬晓飞, 赵东阳. 人体检测与异常行为识别联合算法[J]. 科学技术与工程, 2023, 23(8): 3370-3378.
Ji Xiaofei, Zhao Dongyang. Joint algorithm for human detection and abnormal behavior recognition[J]. Science Technology and Engineering, 2023, 23(8): 3370-3378.

[14] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems. New York: IEEE, 2014: 568-576.

[15] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//European Conference on Computer Vision. Cham: Springer, 2016: 20-36.

[16] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. arXiv Preprint, 2012: arXiv:1212.0402.

[17] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[C]//2011 International Conference on Computer Vision. New York: IEEE, 2011: 2556-2563.

[18] Demir U, Rawat Y S, Shah M. Tinyvirat: low-resolution video action recognition[C]//25th International Conference on Pattern Recognition (ICPR). New York: IEEE, 2021: 7387-7394.

[19] 王琼, 王旭, 刘云麟, 等. 基于改进密集轨迹算法的人体行为识别[J]. 计算机仿真, 2022, 39(12): 284-289, 356.
Wang Qiong, Wang Xu, Liu Yunlin, et al. Human behavior recognition based on improved dense trajectory algorithm[J]. Computer Simulation, 2022, 39(12): 284-289, 356.

[20] Yang M, Guo Y, Zhou F, et al. TS-D3D: a novel two-stream model for action recognition[C]//International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). New York: IEEE, 2022: 179-182.

[21] Bai S, Wang Q, Li X. MFI: multi-range feature interchange for video action recognition[C]//25th International Conference on Pattern Recognition (ICPR). New York: IEEE, 2021: 6664-6671.

[22] Zolfaghari M, Singh K, Brox T. Eco: efficient convolutional network for online video understanding[C]//Proceedings of the European Conference on Computer Vision(ECCV). New York: IEEE, 2018: 695-712.