文章编号:1671-4598(2021)11-0165-06 DOI:10.16526/j.cnki.11—4762/tp.2021.11.030 中图分类号:TP3 文献标识码:A

# 多模态特征融合的长视频行为识别方法

王 婷,刘光辉,张钰敏,孟月波,徐胜军
(西安建筑科技大学 信息与控制工程学院,西安 710055)

**摘要**:行为识别技术在视频检索具有重要的应用价值;针对基于卷积神经网络的行为识别方法存在的长时序行为识别能力不足、尺度特征提取困难、光照变化及复杂背景干扰等问题,提出一种多模态特征融合的长视频行为识别方法;首先,考虑到长时序行为帧间差距较小,易造成视频帧的冗余,基于此,通过均匀稀疏采样策略完成全视频段的时域建模,在降低视频帧冗余度的前提下实现长时序信息的充分保留;其次,通过多列卷积获取多尺度时空特征,弱化视角变化对视频图像带来的干扰;后引入光流数据信息,通过空间注意力机制引导的特征提取网络获取光流数据的深层次特征,进而利用不同数据模式之间的优势互补,提高网络在不同场景下的准确性和鲁棒性;最后,将获取的多尺度时空特征和光流信息在网络的全连接层进行融合,实现了端到端的长视频行为识别;实验结果表明,所提方法在 UCF101 和 HMDB51 数据集上平均精度分别为 97.2%和 72.8%,优于其他对比方法,实验结果证明了该方法的有效性。

**关键词**:深度学习;行为识别;特征提取;多模态特征融合

# Long Video Action Recognition Method Based on Multimodal Feature Fusion

WANG Ting, LIU Guanghui, ZHANG Yumin, MENG Yuebo, XU Shengjun
(College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)

**Abstract**: Action recognition technology has important application value in video retrieval. In order to solve the problems of convolutional neural network based action recognition methods, such as insufficient ability of long time sequence action recognition, difficulty in scale feature extraction, illumination change and complex background interference, a long-video action recognition method based on multi-mode feature fusion is proposed. Firstly, considering that the gap between the frames of the long-sequence behavior is small, it is easy to cause the redundancy of the video frames. Based on this, the time-domain modeling of the whole video segment is completed by using the uniform sparse sampling strategy, and the long-sequence information is fully retained on the premise of reducing the redundancy of the video frames. Secondly, multi-column convolution is used to obtain multi-scale spatial and temporal features, so as to weaken the interference caused by the change of perspective on video images. Then, the optical flow data information is introduced, and the deep features of the optical flow data are obtained through the feature extraction network guided by the spatial attention mechanism. Furthermore, the complementary advantages among different data modes are utilized to improve the accuracy and robustness of the network in different scenarios. Finally, the obtained multi-scale spatial and temporal features and optical flow information are fused in the full connection layer of the network to realize end-to-end long video action recognition. Experimental results show that the average accuracy of the proposed method on UCF101 and HMDB51 datasets is 97.2% and 72.8%, respectively, which is better than other comparison methods. The experimental results prove the effectiveness of the method.

**Keywords**: deep learning; action recognition; feature extraction; multimodal feature fusion

## 0 引言

随着信息时代的快速发展,网络视频数量日创新高,如果不对视频内容加以检索,视频可能会成为谣言的载体,对社会带来不利影响。传统的视频检索是依靠人进行分析检查,而行为识别技术可以代替人工检索,在大量视频数据库中自动检索出指定的行为类别,为视频筛选检查提供技术支持。

针对视频行为识别问题,研究人员先后提出了各种各样的方法,目前,行为识别方法主要可以分为基于传统机器学习[1-2]、基于深度学习[3]两大类。基于传统机器学习的视频行为方法主要是通过类似背景减除法提取人员整体轮廓[4-5]或者诸如时空兴趣点[6]、Harris 角点[7]等局部特征,但此方法所提取的行为特征单一、特征提取过程复杂且工作量较大,对高遮挡、光照变化、背景等因素较为敏感。随着社会的发展,诸多领域对行为识别任务提出了更高的要求,这些方法受其自身的局限性,已无法满足行为识别任务的精度要求。

基于深度学习的方法因具备获取输入数据隐含的深层

收稿日期:2021-04-08; 修回日期:2021-05-13。
基金项目:陕西省自然科学基金面上项目(2020JM—473,2020JM—472);西安建筑科技大学基础研究基金项目(JC1703);西安建筑科技大学自然科学基金项目(ZR19046)。
作者简介:王 婷(1993-),女,陕西西安人,硕士研究生,主要从事机器视觉、行为识别方向的研究。
引用格式:王 婷,刘光辉,张钰敏,等.多模态特征融合的长视频行为识别方法[J].计算机测量与控制,2021,29(11):165-170,175.

次特征的能力，在图像分类、场景分割、文本识别等领域有着广泛的应用[8-10]，同样被研究人员用于行为识别任务中。Simonyan 等人[11]首次提出双流网络模型，该模型包含两个卷积神经网络分支，通过从视频图像中获取不同的输入数据模态进行特征提取，进而提取视频数据的空间信息和时间信息。但因其仅通过一帧来解决空间建模问题，对视频的时域建模能力十分有限[12]。为了解决这一问题，文献［13］提出一种时域分割网络（TSN，temporal segment network），通过对视频进行时域分割和稀疏采样，从输入视频的多个时域片段中随机抽取一个片段，最后聚合不同片段的输出信息得到视频级识别结果，但该网络在特征提取过程中忽略了视频帧在时间维度上的动态相关性，且网络分支过多，不适用于长视频预测。

因长视频预测主要的解决思路是获取视频数据的时空信息，针对于此问题，文献［14］使用长短时记忆网络（LSTM，long short term memory network）分别获取视频数据的全局信息和局部关键信息，其中第一个 LSTM 网络对输入图像包含的完整骨架信息进行编码，获取视频的全局特征信息并从中选择出包含较多信息的关键点，第二个 LSTM 网络对信息量大的关节进行特征提取获得局部关键信息。该方法能够利用 LSTM 网络对视频帧序列进行处理，进而获取长视频的时域信息，但输入数据为骨架序列，限制了其在多类别行为识别任务中的应用。

文献［15］将卷积操作扩展至时间维度，提出 3D 卷积结构，解决了视频识别任务中时空特征提取的问题。文献［16］提出一种基于 3D 卷积的特征提取网络（C3D，convolutional 3 dimention），该网络通过将 3D 卷积核设置为 3×3×3，使其能够提取有效且紧凑的时空特征，在许多视频级任务中获得了较好的结果。然而，在真实场景下，摄像机的高度和角度并不相同，导致拍摄的视频图像在视角上呈现较大差异，C3D 网络不适合用于处理尺度特征变化较大的视频数据。为进一步提升 3D 卷积网络的识别性能，文献［17］提出长时序卷积结构（LTC，long－term temporal convolutions），该方法通过改变网络的输入数据量，使其能够在不同时长的视频数据中保持良好的行为识别性能，但网络参数需随视频长度动态变化，使得该方法难以在实际任务中得到应用。另外，由于 3D 卷积操作在进行时空特征提取时面向的是视频图像帧，因而当视频图像本身受光照变化、复杂背景等因素干扰较多时，网络所提取的时空特征难以对视频中的行为进行有效表征，可能会导致错误的预测。

由上述可知，3D 卷积在一定程度上解决了视频级任务中时空特征提取的问题，但上述方法仍无法有效解决长时域建模能力不足、视角变化导致的全局时空特征提取能力差、光照变化和复杂背景干扰等问题，基于此，本文提出一种多模态特征融合的长视频行为识别方法（long video action recognition method based on multimodal feature fusion）。首先，在数据采样阶段建立整个视频段的时域建模；其次，通过不同大小的 3D 卷积核获取多尺度时空特征，弱化视角变化对视频图像带来的干扰；后引入光流数据信息，通过空间注意力机制引导的特征提取网络获取光流数据的深层次特征，通过不同数据模式之间的优势互补，提高网络在不同场景下的准确性和鲁棒性；最后，将获取的多尺度时空特征和光流信息在网络的全连接层进行融合，实现端到端的长视频行为识别。

## 1 基于多模态特征融合的长视频行为识别方法

本文整体技术路线如图 1 所示，首先，以长视频数据为处理对象，以全视频时域建模为出发点，基于多列卷积的特征提取网络提取能够适应于视角变化的全局时空特征，基于注意力机制引导的特征提取网络获取光流数据的深层次特征；而后，在全连接层进行特征融合并利用 Softmax 分类器完成最终行为识别。

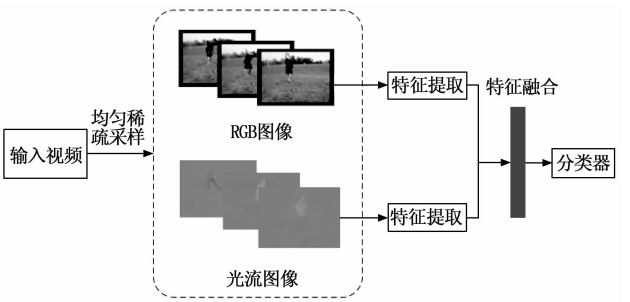


图 1 长视频行为识别算法框架图

### 1.1 视频采样

本文提出的网络主要为识别长视频，但考虑到长时序行为帧间差距较小，随机采样容易引入大量冗余信息，并消除视频图像在时间维度上的相关性，因此在获取全视频段的长时时域信息，建立视频级特征提取网络时，本文引入均匀稀疏采样策略完成全视频段的时域建模，在降低视频帧冗余度的前提下实现长时序信息的充分保留。假设当前视频剪辑有 $N$ 张特征图，则当前采样值 $S$ 可以表示为：

$$S=[N/l] \tag{1}$$

式中，$S$ 代表当前视频的采样值，$N$ 代表当前视频经数据预处理后的图像帧数，$l$ 代表网络输入的数据量，根据得到的采样值 $S$ 对特征图进行位置索引，得到模型输入 $L=[L_0,L_S,\ldots L_{(l-1)S},L_{lS}]$。

本文提出的采样方法类似于 LTC，同样需要计算每个视频的时长，但与 LTC 不同的是，LTC 是根据视频的时长改变网络输入的数据量和输入图像的分辨率，本文通过动态采样值保证了网络输入数据量的一致性，无需调整其余参数，能够适用于不同时长的视频数据。

### 1.2 多尺度时空特征提取网络结构

由于卷积神经网络在逐层提取特征时，输入图像会随着池化操作逐层降低图像分辨率。以往用于行为识别的 3D 卷积神经网络没有考虑低层特征对于时空特征向量生成的影响，而行为识别任务不仅仅关注于运动主体本身的动作，与场景的空间信息也存在密切关系，基于此，本文设计了

一种多尺度时空特征提取网络，具体结构如图 2 所示。网络主要包括 3 部分：多尺度卷积模块、基础骨架网络(C3D)、多特征信息聚合。首先，通过多尺度卷积模块获取原始图像的全局特征，而后利用基础骨架网络生成高低层时空特征，最终通过多特征聚合模块的语义特征嵌入融合方式，将高层时空特征包含较多的语义信息引入低层时空特征，增强低层时空特征的语义表达，使得上下文时空信息和尺度信息相互补充，提高网络对时空特征的表征能力。

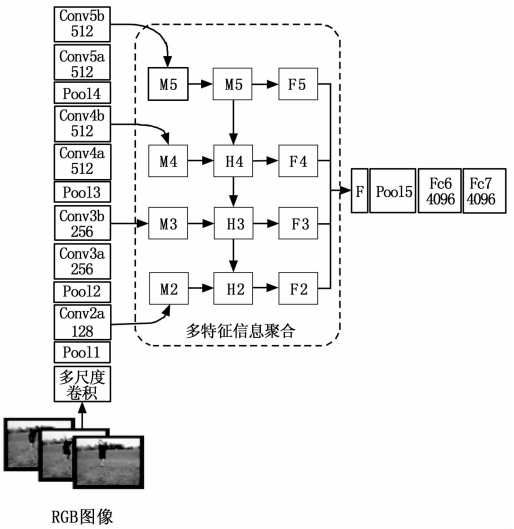


图 2　多尺度时空特征提取网络

### 1.2.1　多尺度卷积

由于拍摄视频时往往存在视角的动态切换，导致视频图像存在较大的尺度变化，而单列卷积难以应对视频图像中的尺度变化问题。因此，本文设计了一种用于时空特征提取的基于多列结构的多尺度卷积模块，具体结构如图 3 所示。在多尺度卷积模块中，采用 3 个不同大小的 3D 卷积核从原始的输入图像块中学习与尺度相关的特征，实现多尺度信息的有效获取，本文采用的多尺度卷积块结构如图 4 所示，经实验验证，采用 3×3×3、5×5×5、7×7×7 的卷积核能够有效聚合全局时空信息。

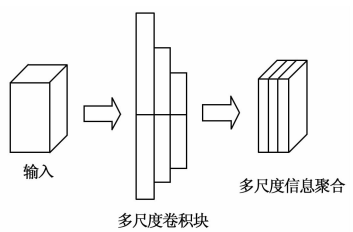


图 3　多尺度模块卷积结构

### 1.2.2　基础骨架网络

本文采用基础骨架网络（C3D）进行特征提取，该网络以堆叠的视频 RGB 帧作为输入数据，再利用 3D 卷积核进行特征提取，卷积核大小决定了提取视频特征的有效性，

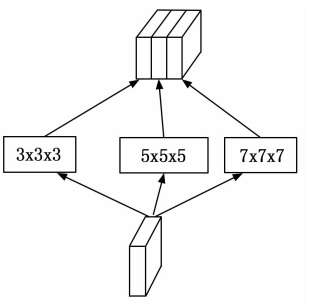


图 4　多尺度卷积块结构

由于视频图像存在动态遮挡、视角变化等问题，这就要求网络所提取的特征必须是通用而有效的，同时在时间维度上，视频特征之间的联系要紧凑，基于此，C3D 网络包含的 8 个 3D 卷积层中所有的卷积核大小均被设置为 3×3×3；池化层均采用最大池化操作，其中，pool1 内核为 1×2×2，其余池化内核均为 2×2×2；网络共有 2 个全连接层，主要用于对特征向量进行降维。网络结构如图 5 所示。为了应对视频图像中的尺度变化问题，本文将该网络的第一个卷积层替换为多尺度卷积模块，通过多尺度卷积获取原始图像的全局特征。

### 1.2.3　多特征信息聚合

在多尺度时空特征提取过程中，3D 卷积核在 $x$，$y$，$z$ 方向上同时移动，神经网络第 $i$ 层第 $j$ 个特征图 $V_{ij}^{xyz}$ 在($x$，$y$，$z$) 处的计算过程如下：

$$V_{ij}^{xyz}=f\left(\sum_{m}\sum_{l=0}^{L_i-1}\sum_{w=0}^{W_i-1}\sum_{h=0}^{H_i-1}W_{i,j,m}^{lwh}\cdot V_{i-1,m}^{(x+l)(y+w)(z+h)}+b_{i,j}\right) \quad (2)$$

式中，$m$ 表示第 $i-1$ 层中与当前特征图相连的特征图；$L_i$ 与 $W_i$ 表示卷积核的长度和宽度；$H_i$ 表示卷积核在时间维度上的尺寸；$W$ 代表与 $i-1$ 层相连的第 $m$ 个特征图的连接权值；$b_{i,j}$ 表示第 $i$ 层第 $j$ 个特征图的偏置；$f$ 为 ReLu 激活函数。

随着卷积层网络的加深，卷积过程会丢失一部分特征信息，由于高层时空特征网络的感受野比较大，所提取的高层时空特征中包含的语义信息较多，空间细节特征较少；低层时空特征网络的感受野比较小，所提取的低层时空特征中包含的空间细节信息较多，高级语义信息较少，如果缺失高层语义信息或低层空间细节信息，均会影响最终的行为识别结果，导致精度降低。针对这一问题，本文构建了一个多特征信息聚合模块，用于聚合高低层时空特征。首先，利用 4 个并行的 1×1×1 卷积核将高低层时空特征的通道值均设置为 512；然后，通过语义嵌入的方式，对高层特征重采样与次高层特征进行自顶向下融合，将高层语义信息用于改进低层的细节信息，再对融合后的特征进行重采样与下一层特征进行融合，增强低层时空特征的语义表达。本文采用的时空特征语义嵌入融合算法如下：

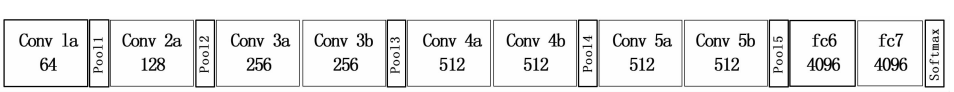


图 5　C3D 网络结构图

$$H_l = Upsample(M_{l+1}) + M_l \tag{3}$$

式中，$H_l$ 表示在 $L$ 层语义嵌入后的时空特征；$M_{l+1}$、$M_l$ 分别为通道值为 512 的高低层时空特征。

之后，采用不同步长的 3×3×3 卷积核将时空特征图映射为具有相同维度的特征图；最后，将嵌入语义信息后的高低层时空特征进行融合，融合后的高低层时空特征 $F_{hl}$ 计算公式如下所示：

$$F_{hl} = \sum_{l_{\min}}^{l_{\max}} F_l \tag{4}$$

式中，$F_l$ 表示在 $L$ 层的时空特征；$l_{\max}$、$l_{\min}$ 分别为最高层及最低层特征索引位置。

### 1.3 光流特征提取

真实场景下视频图像容易受视角和光照变化、复杂背景等因素干扰，因此仅将视频帧作为网络的输入模态难以对视频中的行为进行有效表征，鉴于此，本文引入光流数据（Optial Flow）作为模型的又一输入模态，采用光流信息的原因主要在于：①光流是空间运动物体在观测平面上像素运动的瞬时速度，能够反映视频图像中运动主体的速度、方向等信息；②光流具有表观不变性，表现在视频中的复杂背景及运动主体本身差异性不会影响光流的表现形式[18]。

基于此，本文设计了光流特征提取网络，具体结构如图 6 所示。将光流图作为网络的又一输入模态，以减少光照变化、复杂背景等因素的干扰。以往用于提取光流特征的网络结构较浅，对光流信息的提取更关注于浅层细节信息，而忽略了光流中更深层次的高级语义信息，为充分挖掘光流数据的潜在特征，使用深度残差网络[19] ResNet101 模型作为基础结构，考虑到光流图中的关键信息往往聚集在动作发生的区域，本文在基础网络中添加了空间注意力机制，通过空间注意力选出关键信息，再送入残差网络进行特征提取。

注意力机制的本质就是定位到与当前任务相关的区域[20]，抑制无关信息。由于光流图呈现的内容是动作发生显著变化的区域，所以通过空间注意力机制能够有效定位到图像中的关键信息，有效提升网络性能。

本文采用的空间注意力模型完整结构如图 7 所示，对于特征映射 $F$，首先经过一个最大池化层和一个平均池化层获得两个大小为 $1 \times H \times W$ 特征图，再通过一个 7×7 大小的卷积层获得点对点的空间信息，然后使用 *sigmoid* 函数对空间信息进行激活，得到最终得到的空间注意力激活图 $M_s$，具体如公式（5）所示。

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F);MaxPool(F)])) = \sigma(f^{7\times7}([F^s_{avg};F^s_{\max}])) \tag{5}$$

式中，$\sigma$ 表示 *sigmoid* 函数；$f^{7\times7}$ 表示卷积核大小为 7×7 的卷积运算；$F^s_{avg} \in R^{1\times H\times W}$、$F^s_{\max} \in R^{1\times H\times W}$ 分别为最大池化层和平均池化层输出的特征图。

### 1.4 多模态特征融合

不同模态的特征向量可以通过简单的 Add 相加进行融合，或是在网络末端的决策层进行简单的得分融合，但这些融合方法忽略了特征向量间可能存在的语义冲突关系，导致多模态特征融合过程中可能出现语义信息弱化现象，从而导致模型精度降低。基于此，本文通过构建并训练基

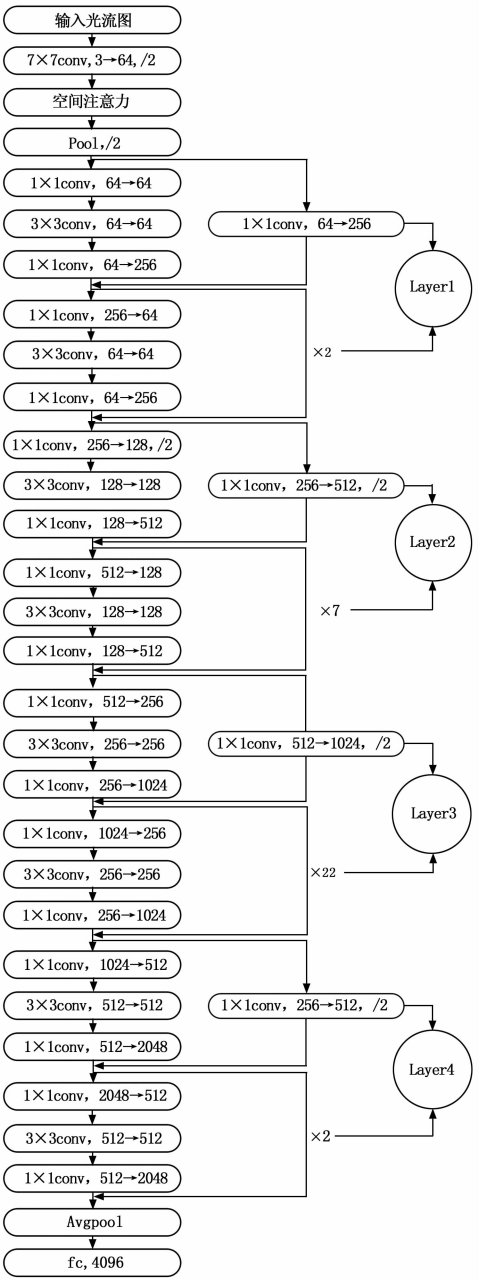


图 6 光流特征提取网络结构图

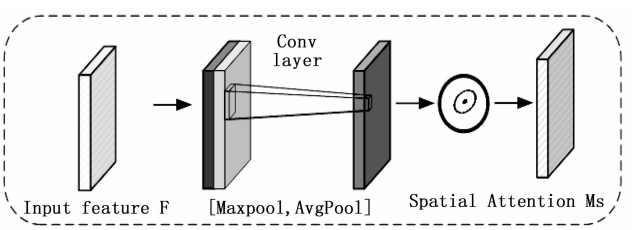


图 7 空间注意力模块

于全连接层的多模态特征融合网络结构，将 4096 维度的多尺度时空特征和 4096 维度的光流数据特征映射到 4096 维的特征融合空间，这种特征融合方式的优势主要在于模型能够在训练阶段学习两个并行网络各自的特征参数，并自主完成协调反馈，实现了模型的端到端训练。

### 1.5　分类器

人体行为的多样性、复杂性要求在进行任务分类时必须更多地保留特征的有用信息，因此，本文选择 Softmax 函数将特征向量映射成概率序列，以保留更多特征的原始信息。Softmax 计算输出类别 $y^{(i)}$ 的过程如公式（6）所示。

$$P(y^{(i)}=k)=\frac{\exp(\eta^j)}{\sum_{j=1}^{k}\exp(\eta^j)},y\in[1,k] \tag{6}$$

式中，$\eta^j$ 为融合后的特征值；$k$ 为类别数；$P$ 表示 $y^{(i)}$ 属于类别 $k$ 的概率值。

## 2　实验与结果分析

本文及对比算法均在 Ubuntu16.04 系统下进行，GPU 型号为 RTX 2080Ti，实验环境配置为 CUDA10.2＋anaconda3＋python3.7＋ pytorch－1.12.0。模型训练过程采用小批量随机梯度下降算法，网络初始训练学习率为 1e－3，迭代次数为 500 次。此外，为使模型充分训练，本文采用数据增强方法，对样本图像进行随机裁剪、旋转、放缩等操作，增强网络模型的鲁棒性。

### 2.1　实验数据集

本文在 UCF101[21] 数据集及 HMDB51[22] 数据集上进行了实验与实验结果分析，其中 UCF101 数据集是从 YouTube 视频网站收集的人类日常活动的视频，共计 13 320 个视频数据，包括 101 个动作类别；HMDB51 数据集大部分数据来源于互联网和电影剪辑，视频图像受光照和视角变化、背景遮挡等因素影响较大，共计 6 849 个视频数据，包括 51 个动作类别。实验按照 UCF101 和 HMDB51 数据集官方给出的 3 种原始划分方案进行训练和测试，并以 3 种划分方案的平均准确率作为最终识别结果。为了便于训练，对数据集进行视频帧截取和光流提取，其中，视频帧通过 ffmpeg 截取，光流图通过 dense _ flow 工具提取，预处理后的数据如图 8 所示，从左到右依次为 RGB 图、$X$ 方向光流图、$Y$ 方向光流图。其中，RGB 图像素大小为 128×171，光流图像素大小为 240×320。

图 8　RGB 图、$x$ 方向光流图及 $y$ 方向光流图

### 2.2　评价指标

为综合评价模型的分类性能，采用准确率（Accuary）作为衡量模型的评估指标，即：

$$Accuary=\frac{TP+TN}{TP+FP+FN+TN}\times 100\% \tag{7}$$

式中，$TP$ 和 $TN$ 表示被正确分类的样本数据；$FP$ 和 $FN$ 表示被错误分类的样本数据。

### 2.3　实验结果定性分析

表 1　本文方法在不同输入模态下的实验结果　（%）

| | 模态 | UCF101 | HMDB51 |
|---|---|---|---|
| 单模态 | 仅基于 RGB 帧 | 89.1 | 50.6 |
| | 仅基于光流 | 93.4 | 64.7 |
| 多模态 | RGB 帧＋光流 | 97.2 | 72.8 |

分析表 1 可知：（1）在基于单模态输入的视频行为识别任务中，仅基于光流的行为识别方法的识别准确度要高于仅基于视频 RGB 帧的方法，在 HMDB51 数据集上，视频图像受光照变化、背景遮挡等因素影响较大，仅基于光流的行为识别方法的识别准确度比仅基于视频 RGB 帧的方法高出 14.1%，说明相较于视频图像，光流具有更强的特征贡献率；（2）多模态融合的行为识别方法的准确度要高于单一模态输入的行为识别方法，说明在行为识别任务中，多模态融合的方法能够结合不同数据模式的优势互补，有效提升行为识别精度。

### 2.4　不同采样方式对比

表 2 显示了本文设计的网络结构在 UCF101、HMDB51 数据集上使用不同采样方式的实验结果，其中 SI 表示本文所提出的均匀稀疏采样，RI 表示随机采样，RSI 表示消除时间相关性后的 SI 采样数据。

表 2　不同采样方式在 UCF101 和 HMDB51 数据集上的实验结果　%

| 采样方式 | UCF101 | HMDB51 |
|---|---|---|
| RI | 91.6 | 65.1 |
| RSI | 93.6 | 69.6 |
| SI | 97.2 | 72.8 |

实验结果表明，本文所用的均匀稀疏采样策略比随机采样具有更高的识别准确率，原因在于随机采样引入了大量的冗余信息；消除采样数据的时间相关性后，时空特征网络和光流信息网络的识别准确率均有所下降，说明了时间维度信息在视频行为识别中的重要性。

### 2.5　与其他方法对比

为综合验证本文所提方法的有效性，本文将单输入模态下的实验结果与当前主流的行为识别方法进行对比，具体如表 3 和表 4 所示。

分析表 3 可知：（1）较之对比算法 LTC，本文设计的多尺度特征提取网络在 UCF101 和 HMDB51 数据集上分别提高了 6.7%和 0.9%，验证了多尺度特征提取网络的有效性；（2）HMDB51 数据集上的识别准确度均偏低，说明当视频图像受光照变化和背景遮挡等因素影响较大时，仅以

视频 RGB 帧作为网络输入数据模态的方法具有一定的局限性。

表 3 基于 RGB 单模态的行为识别方法在 UCF101 和 HMDB51 数据集结果比较 (%)

| 方法 | UCF101 | HMDB51 |
|---|---|---|
| Spatial Stream[11] | 73.0 | 40.5 |
| C3D (1net)[16] | 82.3 | — |
| C3D (3net)[16] | 85.2 | — |
| $LTC_{RGB}$[17] | 82.4 | 49.7 |
| Ours(RGB) | 89.1 | 50.6 |

表 4 基于 Optial Flow 单模态的行为识别方法在 UCF101 和 HMDB51 数据集结果比较 (%)

| 方法 | UCF101 | HMDB51 |
|---|---|---|
| Temporal Stream[11] | 83.7 | 54.6 |
| $TSN_{Flow}$[13] | 87.9 | — |
| $LTC_{Flow}$[17] | 82.4 | 49.7 |
| Ours(Flow) | 93.4 | 64.7 |

分析表 4 可知，本文设计的光流特征提取网络能够有效获取光流数据的深层次特征，较之其它对比算法，对光流特征图的特征分类性能有明显提升。

为验证本文所提方法的有效性，将最终实验结果与当前主流方法进行了比较，具体如表 5 所示。

表 5 UCF101 和 HMDB51 数据集上本文方法与其他算法比较 (%)

| 方法 | UCF101 | HMDB51 |
|---|---|---|
| DT+MVSV[1] | 83.5 | 55.9 |
| IDT+FV[2] | 85.9 | 57.2 |
| Two-stream (avg. fusion)[11] | 86.9 | 58.0 |
| Two-stream (SVM fusion)[11] | 88.0 | 59.4 |
| Two-stream fusion[12] | 88.6 | — |
| TSN[13] | 94.0 | 69.2 |
| C3D+IDT[16] | 90.4 | — |
| $LTC_{RGB+Flow}$[17] | 91.7 | 64.8 |
| LTC+IDT[17] | 92.7 | 67.2 |
| Ours | 97.2 | 72.8 |

分析表 5 可知，本文所提方法在 UCF101 和 HMDB51 数据集上有良好的表现，较之对比算法，不仅能够识别对长视频中的人体行为，且具有更高的识别准确率。

## 3 结束语

本文提出一种多模态特征融合的长视频行为识别方法，网络首先在数据采样阶段引入了均匀稀疏采样策略，进而完成全视频段的时域建模，其次，通过多列卷积获取多尺度时空特征，弱化视角变化对视频图像带来的干扰，后引入光流数据信息，通过空间注意力机制引导的特征提取网络获取光流数据的深层次特征；最后，将获取的多尺度时空特征和光流信息在网络的全连接层进行融合，实现了端到端的长视频行为识别，解决了基于卷积神经网络的视频行为识别方法存在的长时序行为识别能力不足、尺度特征提取困难、光照变化及复杂背景干扰等问题。在 UCF101 和 HMDB51 数据集上的实验结果验证了本文方法的有效性。

**参考文献：**


[1] WANG H, KLASER A, SCHMID C, et al. Action recognition by dense trajectories [C] //Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition. Jun 11, 2011, Colorado Springs, United States. New York: IEEE, 2011: 3169-3176.

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

[3] 蔡 强, 邓毅彪, 李海生, 等. 基于深度学习的人体行为识别方法综述 [J]. 计算机科学, 2020, 47 (4): 85-93.

[4] BOBICK A F, DAVIS J W. The recognition of human movement using temporal templates [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23 (3): 257-267.

[5] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as space-time shapes [J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29 (12): 2247.

[6] KONSTANTINOS R, YANNIS A, STEFANOS K. Dense saliency-based spatiotemporal feature points for action recognition [C] //2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009.

[7] LAPTEV I. On Space-Time Interest Points [J]. International Journal of Computer Vision, 2005, 64 (2/3): 107-123.

[8] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39 (6): 1137-1149.

[9] CAO Y L, MING T F, HE G, et al. Artificial Recognition of centrifugal pump cavitation status based on deep learning [J]. Journal of Xi'an Jiaotong University, 2017, 51 (11): 165-172.

[10] CHANG L, DENG X M, ZHOU M Q, et al. Convolutional Neural Networks in Image Understanding [J]. Journal of Automation, 2016, 42 (9): 1300-1312.

[11] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [C] // Proceedings of the 2014 Annual Conference on Neural Information Processing Systems, Montreal, Dec 8-13, 2014. Red Hook: Curran Associates, 2014: 568-576.


(下转第 175 页)