基于特征交互和聚类的行为识别方法

李凯歌, 蔡鹏飞, 周忠*

(北京航空航天大学虚拟现实技术与系统国家重点实验室 北京 100191)
(zz@buaa.edu.cn)

摘 要: 针对现有行为识别方法缺乏对时空特征关系建模的问题, 提出一种基于特征交互和聚类的行为识别方法. 首先设计一种混合多尺度特征提取网络提取连续帧的时间和空间特征; 然后基于Non-local操作设计一种特征交互模块实现时空特征的交互; 最后基于三元组损失函数设计一种难样本选择策略来训练识别网络, 实现时空特征的聚类, 提高特征的鲁棒性和判别性. 实验结果表明, 与基线方法 TSN 相比, 所提方法的准确度在 UCF101 数据集上提高了 23.25 个百分点, 达到 94.82%; 在 HMDB51 数据集上提高了 20.27 个百分点, 达到 44.03%.

关键词: 行为识别; 时空特征关系; 特征交互; 特征聚类
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023.19493

Action Recognition Based on Feature Interaction and Clustering

Li Kaige, Cai Pengfei, and Zhou Zhong*

(*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing* 100191)

Abstract: To mitigate the problem that the action recognition methods lack the modeling of spatiotemporal feature relationship, an action recognition method based on feature interaction and clustering is proposed. Firstly, a mixed multi-scale feature extraction network is designed to extract spatial and temporal features of continuous frames. Secondly, a feature interaction module is designed based on non-local operation to realize spatiotemporal feature interaction. Finally, based on the triplet loss function, a hard sample selection strategy is designed to train the recognition network, thus realizing spatiotemporal feature clustering and improving the robustness and discrimination of the features. Experimental results show that compared with TSN, the accuracy of on the UCF101 dataset is increased by 23.25 percentage points to 94.82%. On the HMDB51 dataset, the accuracy is increased by 20.27 percentage points to 44.03%.

Key words: action recognition; spatiotemporal feature relationship; feature interaction; feature clustering

随着通信技术、计算机技术以及图像处理技术的飞速发展, 视频监控系统越来越广泛地应用于交通、医疗、娱乐和公共安全等众多领域. 然而, 视频监控系统通常由工作人员负责观看和分析, 不仅耗费了大量的人力资源, 而且难以满足海量视频数据的处理需求. 智能化视频分析技术可以有效地辅助工作人员对海量的视频数据进行分析和处理, 受到了学术界的广泛关注.

行为分析是视频分析的重要内容, 需要从相机的视频数据中分离出含有人的行为信息的视频帧, 并对这些视频帧所包含的行为进行分类. 近年来, 行为分析已经被广泛地应用于人机交互、医疗

收稿日期: 2021-11-10; 修回日期: 2021-12-22. **基金项目**: 国家自然科学基金(61872024); 国家重点研发计划(2018YFB2100603).
李凯歌(1994—), 男, 博士研究生, 主要研究方向为计算机视觉; **蔡鹏飞**(1995—), 男, 硕士研究生, 主要研究方向为计算机视觉; **周忠**(1978—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 论文通信作者, 主要研究方向为虚拟现实、计算机视觉.

看护以及视频安防监控等领域[1]. 随着深度学习, 特别是卷积神经网络(convolutional neural network, CNN)在计算机视觉领域的广泛应用, 行为识别方法的精度得到了显著提升.

行为指发生在一段时间内、具有特定空间模式的一组动作, 因此, 表观信息和运动信息的捕捉对于行为识别具有重要意义. 然而, 目前用于行为识别的双流网络通常单独训练 RGB 流(表观)和光流(运动)特征提取分支, 未考虑两者之间的互补关系. 另外, 真实视频中场景和行为主体的多样性造成了同类行为表观差异大、不同行为表观相似的问题, 现有行为识别方法中的特征提取网络无法有效地捕捉视频帧中的不同尺度信息, 导致其性能很难继续提高.

针对上述问题, 本文提出一种基于多尺度特征交互和聚类的行为识别方法. 首先, 通过设计混合多尺度特征提取网络(mixed multi-scale network, MMSNet)来提取适用于行为识别的多尺度时空特征, 应对复杂场景下的行为表观变化; 其次, 通过设计基于 Non-local 操作[2]的特征交互模块(feature interaction module, FIM)来提取时空特征中的互补信息, 实现 2 种模态特征的交互; 再通过特征共享模块和基于三元组损失函数(triplet loss function, TLF)[3]的样本选择策略来实现时空特征在高维嵌入空间上的聚类; 最后, 将基于 2 种模态特征的分类得分进行融合, 得到最终的行为分类结果. 实验结果表明, 本文方法显著地改进了时序分段网络(temporal segment network, TSN)[4]的性能, 可以实现更加鲁棒、准确的行为识别.

1 相关工作

行为是时间和空间上的一种连续变化, 即时间(表观)信息和空间(运动)信息的融合, 包含着丰富的语义信息. 提取和利用视频中的时空特征是视频行为识别的关键. 基于传统特征的行为识别方法最早被提出[5-6], 但是手工设计的特征易受光照、遮挡和噪声的影响, 导致该类方法的鲁棒性和准确度不高. 近年来, 深度学习方法在图像领域得到广泛应用. 其可以很好地捕获图像的高层语义特征, 在复杂环境下表现出较强的鲁棒性, 已被广泛地应用于行为识别任务. 基于深度学习的行为识别方法分为 4 种类型.

(1) 基于时间序列模型的循环神经网络. Donahue 等[7]提出一种基于长短时记忆网络的长时循环卷积网络(long-term recurrent convolutional networks, LRCN), 通过 CNN 从视频帧中提取表观特征, 然后利用长短时记忆网络对行为种类进行预测. LRCN 中每一帧的特征都对后一帧的特征有所影响, 通过这种方式模拟持续的表观变化表现出的运动信息, 最后对所有的预测值取平均得到行为分类结果. 该类方法由于通常使用时间特征池化来组合时间信息, 因此不能很好地捕获整个视频时间维度上的表示.

(2) 基于 3D 卷积[8]隐式地对时空关系进行建模的神经网络. 使用连续的多帧图像作为输入, 直接应用 2D 卷积会得到扁平的输出, 造成运动信息的损失. Tran 等[9]提出基于 3D 卷积的行为识别网络 C3D, 利用 3D 卷积来保持输入的时间维度信息特性. 3D 卷积由于在一次卷积操作中同时提取表观特征和运动特征, 因此无法很好地对时空信息进行建模. Tran 等[10]在 3D 卷积的基础上提出了 R(2+1)D 网络, 将 3D 卷积操作分解为 2D 卷积和 1D 卷积, 分别对输入进行空间建模和时间建模, 以提高行为识别的精度. Chang 等[11]提出一种长时视频行为识别方法(long-term video action recognition, LVAR), 在 C3D 中引入递归连接来传播时空信息. 然而, 3D 卷积隐式地建模时空关系, 且网络参数普遍高于 2D CNN, 使此类方法的性能难以持续提高.

(3) 基于图卷积的行为识别网络. 相比于包含表观信息的 RGB 帧和包含运动信息的光流帧, 动态的人体骨骼通常可以传达更重要的行为信息. Yan 等[12]首次提出一种基于图卷积的行为识别网络结构, 将视频帧中人的骨架数据作为输入, 使用图卷积完成行为识别. Shi 等[13]针对图卷积拓扑结构固定的问题, 提出一种双流的图 CNN, 分别对骨骼信息和关键点信息进行学习, 以加强行为识别的准确度. 该类方法由于通常采用外接的姿态估计网络所预测出的人体骨架数据作为输入, 因此无法实现端到端的行为识别, 且在复杂场景中可能无法获得整个人体的结构.

(4) 显式地对时空特征进行建模的双流识别网络架构. Simonyan 等[14]提出的 Two-stream 是最早用于行为识别的双流架构, 其性能远超基于传统特征的行为识别方法. Wang 等[4]针对 Two-stream 网络不能考虑整段视频行为信息的缺点, 提出通过分段采样视频帧同时使用 RGB 流分支和光流分支对行为进行分类, 最后对 2 个分支的分类得分进行融合的 TSN. Feichtenhofer 等[15]提出一种双流网络融合方法(two-stream network fusion, TSF), 通过探索时

空特征图的不同融合方式达到了先进的识别性能. 基于双流结构的行为识别网络, 通过分别提取 RGB 流和光流特征的方式显式地对时空信息建模, 不需要利用第三方网络, 且双流结构在行为识别任务上的性能普遍优于基于 3D CNN 和循环神经网络的方法, 因此在视频行为识别任务中得到了广泛应用. 然而, 现有的此类方法由于通常单独提取表观特征和运动特征, 忽略了两者之间的互补关系, 因此网络的判别能力有待进一步提高. 如图 1 所示, 喝水和进食 2 种行为的光流(运动)特征极为相似, 但基于 RGB(表观)特征可以很好地对 2 种行为进行分类, 说明了表观特征和运动特征的融合、互补信息对行为识别的结果具有重要作用.

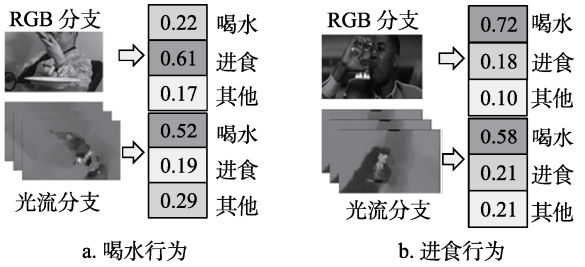


图 1　时空特征之间的互补性示意图

为了同时对时空特征进行学习, Feichtenhofer 等[16]提出了一种基于时空特征乘性交互的视频行为识别方法, 通过基于运动特征的门控机制融合了双流体系中的表观和运动特征, 实现了端到端的训练. Cho 等[17]提出的时空融合网络通过融合 3 个分段的特征形成新的 RGB 流特征和光流特征, 并采用 3 种特征融合策略来提高识别性能, 即平均值融合、最大值融合以及分类得分融合. 然而, 上述 2 种方法采用的是简单、机械的融合策略, 而且单独训练 2 种模态数据的特征提取分支, 忽略了 2 种模态特征之间的互补信息. 另外, 由于 2 种模态特征之间存在差异, 因此此类方法可能会降低原有的分类准确度. 为此, 本文提出了一种可以同步训练双流特征提取分支的行为识别方法, 通过基于 Non-local 操作的 FIM 来充分利用 2 种模态特征的互补信息, 通过共享模块和 TLF, 在维持不同模态特征独有信息的同时学习特征间的共同表示, 以此实现更加鲁棒、精准的行为识别.

2　本文方法

本文提出一种用于行为识别的同步双流行为识别网络, 并通过设计轻量级的 MMSNet、基于 Non-local 操作的 FIM 和基于难样本挖掘 TLF 的时空特征聚类方法来提高网络性能.

2.1　同步双流行为识别网络结构

现有的双流结构行为识别网络以单流的方式分别训练 RGB 流(空间)分支和光流(时间)分支, 只在网络最后对 2 个分支的分类得分进行融合. 这种训练方式不仅难以对时空特征的关系进行建模, 而且会破坏时空特征之间的联系. 为此, 本文提出一种基于特征交互和聚类的同步双流网络(synchronous two-stream network, STN), 以端到端的形式同时训练空间流分支和时间流分支, 充分利用 2 种模态特征所包含的信息进行行为识别. STN 的整体结构示意图如图 2 所示.

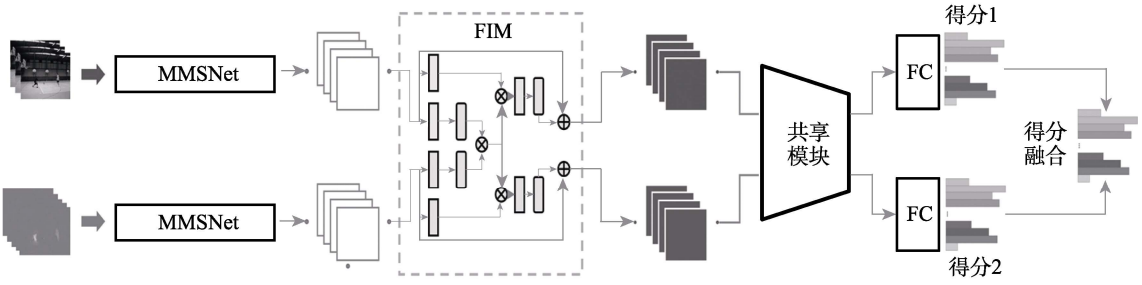


图 2　STN 整体结构示意图

如图 2 所示, 给定 RGB 和光流输入, 首先采用 2 个 MMSNet 分别提取空间特征和时间特征; 然后利用 FIM 加强时空特征之间的交互, 充分利用 2 种模态特征之间的互补信息; 再利用所设计的特征共享模块和 TLF, 使网络在获得 2 种模态信息共同表示的同时又可以使 2 种信息保持自身独有的信息, 实现 2 种模态特征的聚类; 最后对 2 个分支的预测结果进行融合, 进一步增强模型的识别能力.

2.2　MMSNet

现实生活中, 场景变化、行为主体变化或相机视角变化可能会导致相同类别行为具有较大的外观变化(如图 3a 所示), 而不同类别行为的外观极

为相似(如图 3b 所示), 增加了视频行为识别的难度, 对行为分类带来了负面影响. 从图 3a 中可以看出, 2 组同属“滑板”行为的视频片段中, 由于场景和行为个体不同, 同一行为的外观表现出较大差异; 图 3b 中, 2 组视频片段分别代表“扔链球”和“掷铁饼”行为, 虽然相机不同, 但由于场景和视角相似, 2 个行为的表观非常相近, 因此很难区分 2 个行为, 在分类时这 2 个片段可能会被误分为同一行为. 神经网络中的多尺度特征可以有效地解决这个问题, 原因是全局尺度特征可以获得视频帧的全部信息从而缩小类别范围, 并得到行为类别候选. 此时, 可以利用局部特征(如链球或滑板等一些局部区域的特征)对行为进行更加精准的分类. 因此, 为了获得更高的行为识别准确度, 不仅要获取场景的全局特征信息, 局部特征也至关重要, 即多尺度特征的获取是提高行为识别性能的关键.

a. 场景不同而行为相同

b. 场景类似而行为不同

图 3 2 组行为视频帧抽样

现有的行为识别网络通常采用经典的图像分类网络(如 VGG[18], ResNet[19]和 DenseNet[20]等)来提取特征, 而这些单尺度的特征提取网络未考虑视频帧中的多尺度信息, 难以对视频行为进行精准的分类识别. 在场景相似的情况下, 区分不同行为更依赖特征的细微差别, 即小尺度特征; 在场景差异较大的情况下, 区分不同行为往往更依赖大尺度信息. 因此, 本文设计一种由多个多尺度模块堆叠而成的特征提取网络, 每个多尺度模块由 4 个不同尺度的卷积分支和 1 个特征聚合组件构成. 多尺度模块利用具有不同感受野的卷积分支提取不同尺度的特征信息, 通过由共享全连接层(fully connection layer, FC)构成的特征聚合组件实现不同尺度特征的自适应融合. 通过显式地捕获多尺度特征信息, MMSNet 可以显著地提高行为识别的准确度.

2.2.1 多尺度模块

CNN 的感受野表示网络每一层输出特征图上的像素点在输入图像上的映射区域(感受范围)大小. 感受野的大小与卷积层的数量有关, 在堆叠 n 个 3×3 卷积层后, 网络的感受野 R 变为 $2n+1$. Chang 等[21]的实验证明: 网络越浅(感受野小), 越重视浅层细节特征, 如衣服的颜色或纹理; 网络越深(感受野大), 越重视深层语义特征, 如提包的种类或行人的性别等. 因此, 本文可以通过调整卷积层的数量来改变网络的感受野, 从而捕获不同尺度的特征.

为了减少参数量和计算量, 本文采用深度可分离卷积[22]替代标准卷积, 以构建具有残差机制的网络基本残差单元, 并通过堆叠多个残差单元捕获输入特征的多尺度信息. 如图 4 所示, 网络的基础模块——多尺度模块包含 4 个不同尺度的卷积分支, 第 i 个卷积分支由 i 个残差单元堆叠而成, 用于捕获尺度为 $2i+1$ 的特征 $\boldsymbol{F}_i$. 为了以自适应的方式聚合所获得的不同尺度特征, 本文在多尺度模块的尾部添加了 1 个由共享 FC 构成的特征聚合组件. 根据特征的不同, 特征聚合组件为其分配特定的权重来融合不同尺度的特征. 加权融合后的特征为 $\boldsymbol{F}_{\mathrm{f}}=\sum_{i=1}^{m}\left(g\left(\boldsymbol{F}_i\right)+\varepsilon\right)\otimes\boldsymbol{F}_i$. 其中, $g\left(\boldsymbol{F}_i\right)\in\mathbb{R}^{c'}$ 表示共享 FC 输出的不同通道特征的权重; c' 表示输入特征 $\boldsymbol{F}_i$ 的通道数量; ε 表示趋于 0 的常数, 默认为 10^{-5}; $\otimes$ 表示逐通道相乘.

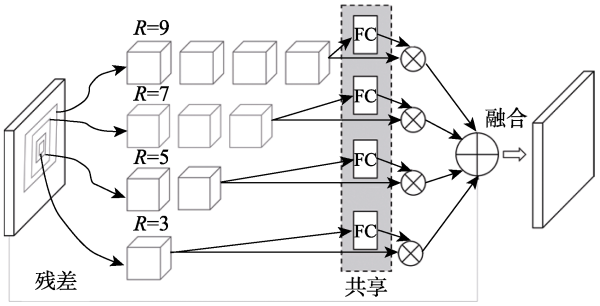


图 4 多尺度模块结构示意图

图 4 中, 每个立方体表示基本残差单元. 本文设计的多尺度模块共有 4 个分支, 用于捕获 4 个尺度的特征信息, 最后通过自适应的特征融合操作获得输出特征. 残差结构用于提高模型的鲁棒性, 以降低梯度消失的风险.

2.2.2 轻量级深度可分离卷积

卷积运算拥有优良的特征提取能力, 比全连

接结构的参数更少, 在图像或视频帧这种二维结构数据上具有显著优势. 为了进一步降低卷积的参数量、加快网络的推理速度, Howard 等[22]提出一种卷积结构——深度可分离卷积, 其核心思想是将标准卷积运算分解为深度卷积和逐点卷积, 如图 5 所示.

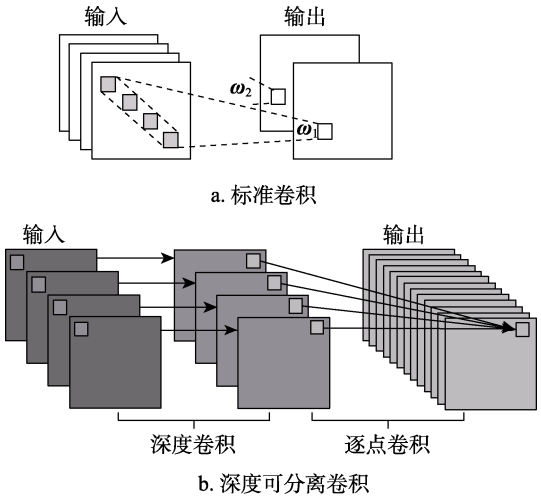


图 5　不同卷积结构对比

假设输入特征图 $\boldsymbol{x}\in\mathbb{R}^{h\times w\times c}$, 标准卷积核 $\boldsymbol{\omega}\in\mathbb{R}^{k\times k\times c\times c'}$. 其中, h 和 w 表示输入特征图的高和宽; c 表示输入特征的通道数; c' 表示输出特征的通道数. 标准卷积操作可以表示为 $\boldsymbol{\omega}*\boldsymbol{x}$, 其参数量为

$$\rho = k^2\times c\times c' \tag{1}$$

深度可分离卷积将标准卷积运算进行分解, 可以表示为 $\left(\boldsymbol{\omega}_p\circ\boldsymbol{\omega}_d\right)*\boldsymbol{x}$. 其中, $\boldsymbol{\omega}_d$ 表示深度卷积的卷积核, 大小为 $\mathbb{R}^{k\times k\times 1\times c}$; $\boldsymbol{\omega}_p$ 表示逐点卷积的卷积核, 大小为 $\mathbb{R}^{1\times 1\times c\times c'}$. 标准卷积的卷积核应用在所有输入通道上. 深度可分离卷积首先采用深度卷积对每个输入通道分别进行卷积, 1 个卷积核对应 1 个通道; 然后采用逐点卷积对深度卷积的输出进行整合, 其整体效果与标准卷积相似, 但会大大减少模型的计算量和参数量. 深度可分离卷积的参数量为

$$\rho = \left(k^2 + c'\right)\times c \tag{2}$$

对比式(1)(2)可以看出, 深度可分离卷积显著地减少了参数量. 另外, 标准卷积的计算量为 $h\times w\times c\times k\times k\times c'$; 而深度可分离卷积的计算量为 $h\times w\times c\times\left(k^2+c'\right)$, 是深度卷积和逐点卷积的计算量之和. 当采用 3×3 深度可分离卷积时, 其计算量比标准卷积减少 1/9. 因此, 基于此种结构, 本文设计了如图 6 所示轻量级的深度可分离卷积结构.

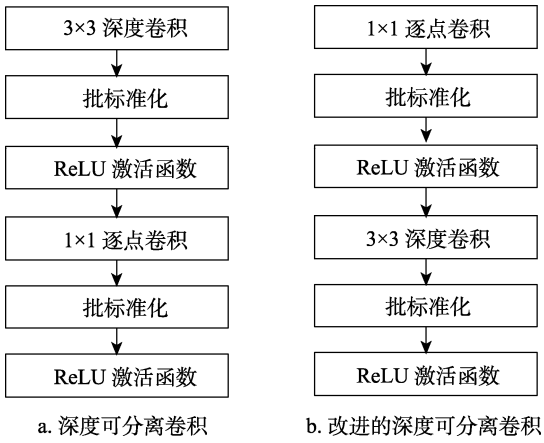


图 6　深度可分离卷积结构对比

与文献[22]结构不同, 本文先进行逐点卷积运算, 再进行深度卷积运算, 这种结构的卷积运算可以表示为 $\left(\boldsymbol{\omega}_d\circ\boldsymbol{\omega}_p\right)*\boldsymbol{x}$. 其中, 逐点卷积核 $\boldsymbol{\omega}_p\in\mathbb{R}^{1\times 1\times c\times c'}$, 深度卷积核 $\boldsymbol{\omega}_d\in\mathbb{R}^{k\times k\times 1\times c'}$. 与原本的深度可分离卷积的区别是, 本文首先使用逐点卷积增加特征通道数量, 提升网络的容量. 然后采用深度卷积提取特征. 此时, 改进的深度可分离卷积的参数量为 $\rho=\left(k^2+c\right)\times c'$.

本文设置输出通道数 $c'>c$, 这会导致参数量相比原本的深度可分离卷积操作有所上升, 但是参数量和计算量仍远小于标准卷积操作. 另外, 实验结果表明, 本文设计的深度可分离卷积在行为识别任务中拥有更加优异的性能.

2.2.3　MMSNet 结构

与仅在网络最后阶段对不同分支特征进行融合的方法(如 TSF[15])不同, MMSNet 由多个可以同时捕获不同尺度特征的多尺度模块堆叠而成. 通过堆叠多尺度模块, 网络可以提取更为鲁棒的混合多尺度特征, 提高网络对不同行为的区分能力, 网络基本残差单元的利用也有效地提高了模型的稳定性. 为了降低模型复杂度和计算量, 本文采用改进的深度可分离卷积构建多尺度模块, 提升网络的训练和推理效率. MMSNet 的结构如图 7 所示.

MMSNet 的输入是采样的视频帧数据或光流帧数据. 以视频帧为例, 当以 3 个视频帧作为输入时, 数据尺寸为 $(b,9,224,224)$. 其中, b 表示输入的批量大小; 每幅图像有 3 个通道, 故输入有 9 个通道; 视频帧的宽和高均为 224. 输入视频帧首先

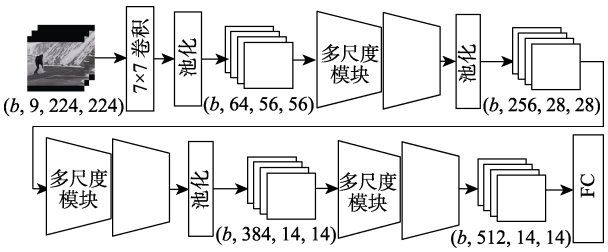


图 7 MMSNet 结构示意图

通过 7×7 卷积层对进行特征编码, 然后通过多个多尺度模块来提取混合多尺度特征, 最后特征通过 FC 得到分类识别结果.

2.3 基于 Non-local 操作的 FIM

由于单独训练空间流和时间流分支会破坏行为的表观和运动特征之间的联系, 因此本文提出的 STN 以 RGB 视频帧和光流帧作为输入, 利用空间流和时间流特征提取分支同时捕获表观特征和运动特征, 然后通过所设计的 FIM 完成 2 种模态特征的交互, 提高特征的鲁棒性并增强模型的识别能力. 其中, 2 个特征提取分支均由 MMSNet 构成, 分别用于提取 2 种模态的特征, 为后续的特征交互和聚类做准备.

假设输入数据中包含 $b\in\mathbb{N}$ 个视频片段(由视频帧序列和光流帧序列 2 部分构成), 表示为 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_b\}$, 对应的标签数据为 $\boldsymbol{L}=\{l_1,l_2,\cdots,l_b\}$. 本文采用稀疏采样, 将每个视频片段分割成 T 段, 每段采样 1 帧 RGB 图像和 β 帧光流图像, 则第 i 个视频片段的第 $t\in[1,\ T]$ 个分段的数据可以表示为 $\boldsymbol{V}_{i,t}=\{r_t,O_t\}$. 其中, r_t 表示第 t 个分段里的 RGB 帧; $\boldsymbol{O}_t$ 表示与 RGB 帧 r_t 对应的 β 幅光流帧, $\boldsymbol{O}_t=\{o_t\}_{j=t-\beta_1}^{j=t+\beta_2}$, 且 $\beta_1+\beta_2=\beta$.

RGB 帧和光流帧分别经过空间流和时间流分支的特征提取网络后, 可以得到 RGB 表观特征 $\boldsymbol{F}^{\mathrm{r}}\in\mathbb{R}^{b\times T\times C\times H\times W}$ 和光流(optical flow)运动特征 $\boldsymbol{F}^{\mathrm{o}}\in\mathbb{R}^{b\times T\times C\times H\times W}$. 此过程表示为 $\boldsymbol{F}_i^{\mathrm{r}},\ \boldsymbol{F}_i^{\mathrm{o}}=\Gamma(\boldsymbol{V}_i)$. 其中, $\Gamma(\cdot)$ 表示混合多尺度特征提取操作; i 表示输入的小批量视频数据里的第 i 个视频片段. 为了便于理解, 本文将 $b\times T$ 记作 B, 即 1 批输入中共有 B 个分段, 此时 $\boldsymbol{F}_i^{\mathrm{r}},\ \boldsymbol{F}_i^{\mathrm{o}}\in\mathbb{R}^{B\times C\times H\times W}$. 在此基础上, 本文设计了一种可充分利用 2 种模态特征互补信息的 FIM, 以增强不同模态特征的判别性.

受图像去噪领域中非局部平均操作[23]和自然语言处理领域中自注意力机制[24]的启发, Wang 等[2]提出一种 Non-local 操作, 用于捕获 CNN 中的长距离依赖关系. 进行特征提取时, Non-local 操作考虑其他所有输入数据的信息来获得输出特征, 该属性使其可以很好地建模特征之间的关系. 为了对时空特征之间的关系进行建模, 本文首先计算两者之间的相似程度, 然后根据此相似度来提取 2 种模态特征中的一致信息, 实现 2 种模态特征的交互. 综上所述, 本文设计的基于 Non-local 操作的 FIM 通过增强时空特征的联系, 使一种模态的特征从另外一种模态的特征中获得互补信息, 从而显著地提高特征的鲁棒性. 特征交互操作公式为

$$\begin{cases}\hat{\boldsymbol{F}}^{\mathrm{r}}=\dfrac{1}{\zeta\left(\boldsymbol{F}^{\mathrm{r}},\boldsymbol{F}^{o}\right)}\displaystyle\sum_{i,j}s\left(\boldsymbol{F}_i^{\mathrm{r}},\boldsymbol{F}_j^{\mathrm{o}}\right)g\left(\boldsymbol{F}_i^{\mathrm{r}}\right)\\ \hat{\boldsymbol{F}}^{\mathrm{o}}=\dfrac{1}{\zeta\left(\boldsymbol{F}^{\mathrm{r}},\boldsymbol{F}^{o}\right)}\displaystyle\sum_{i,j}s\left(\boldsymbol{F}_i^{\mathrm{r}},\boldsymbol{F}_j^{\mathrm{o}}\right)g\left(\boldsymbol{F}_j^{\mathrm{o}}\right)\end{cases}.$$

其中, $i,j\in[1,\ B]$ 表示分段的索引; $\boldsymbol{F}_i^{\mathrm{r}}$ 和 $\boldsymbol{F}_j^{\mathrm{o}}$ 分别表示第 i 和第 j 个分段的表观和运动特征; $\zeta(x,y)$ 表示归一化因子; 函数 $g(\cdot)$ 用于对输入信号进行特征变换; $s(\cdot)$ 表示用于计算特征间相似度的二元函数. 有很多种度量 2 个特征相似程度的方法, 如 Buades 等[23]在图像去噪任务中采用的欧几里得距离度量方法. 与之相比, 点积相似度更加简洁, 也更利于算法实现. 因此, 本文采用在嵌入空间上计算特征点积相似度的方法来度量特征之间的相似程度, 公式为 $s\left(\boldsymbol{F}_i^{\mathrm{r}},\boldsymbol{F}_j^{\mathrm{o}}\right)=\mathrm{e}^{\theta\left(\boldsymbol{F}_i^{\mathrm{r}}\right)^{\mathrm{T}}\phi\left(\boldsymbol{F}_j^{\mathrm{o}}\right)}$. 其中, $\theta\left(\boldsymbol{F}_i^{\mathrm{r}}\right)=\boldsymbol{W}_\theta\boldsymbol{F}_i^{\mathrm{r}}$ 和 $\phi\left(\boldsymbol{F}_j^{\mathrm{o}}\right)=\boldsymbol{W}_\phi\boldsymbol{F}_j^{\mathrm{o}}$ 分别表示 2 种模态特征在嵌入空间上的表示. 由于归一化因子 $\zeta\left(\boldsymbol{F}^{\mathrm{r}},\boldsymbol{F}^{o}\right)=\displaystyle\sum_{i,j}f\left(\boldsymbol{F}_i^{\mathrm{r}},\boldsymbol{F}_j^{\mathrm{o}}\right)$, 因此 FIM 中的 Non-local 操作可公式化表示为

$$\begin{cases}\hat{\boldsymbol{F}}^{\mathrm{r}}=\mathrm{Softmax}\left(\boldsymbol{F}^{\mathrm{rT}}\boldsymbol{W}_\theta^{\mathrm{T}}\boldsymbol{W}_\phi\boldsymbol{F}^{\mathrm{o}}\right)g\left(\boldsymbol{F}^{\mathrm{r}}\right)\\ \hat{\boldsymbol{F}}^{\mathrm{o}}=\mathrm{Softmax}\left(\boldsymbol{F}^{\mathrm{oT}}\boldsymbol{W}_\phi^{\mathrm{T}}\boldsymbol{W}_\theta\boldsymbol{F}^{\mathrm{r}}\right)g\left(\boldsymbol{F}^{\mathrm{o}}\right)\end{cases}.$$

最后, 本文引入残差连接来构建 FIM, 公式为

$$\begin{cases}\boldsymbol{f}^{\mathrm{r}}=\boldsymbol{W}_{\mathrm{r}}\hat{\boldsymbol{F}}^{\mathrm{r}}+\boldsymbol{F}^{\mathrm{r}}\\ \boldsymbol{f}^{\mathrm{o}}=\boldsymbol{W}_{\mathrm{o}}\hat{\boldsymbol{F}}^{\mathrm{o}}+\boldsymbol{F}^{\mathrm{o}}\end{cases}.$$

残差连接使 FIM 可以插入任何预训练的网络模型中, 而不会破坏其初始性能, 提高了 FIM 的鲁棒性和稳定性. FIM 的结构如图 8 所示.

图 8 中, FIM 的输入包括 RGB 流特征 $\boldsymbol{F}^{\mathrm{r}}$ 和光流特征 $\boldsymbol{F}^{\mathrm{o}}$, 尺寸均为 $(B,512,H,W)$. 其中, 512 代表通道数; $\otimes$ 表示矩阵乘法; $\oplus$ 表示矩阵加法; $\boldsymbol{W}_g,\boldsymbol{W}_\theta,\boldsymbol{W}_\phi,\boldsymbol{W}_r,\boldsymbol{W}_o$ 均表示 1×1 卷积. 经过 FIM

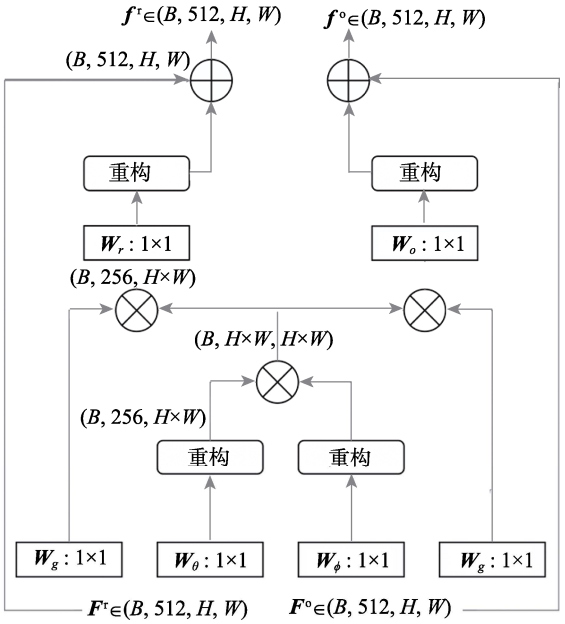


图 8　FIM 结构示意图

后, 输出的特征交互后的 RGB 流特征和光流特征分别为 $\boldsymbol{f}^{\mathrm{r}}$ 和 $\boldsymbol{f}^{\mathrm{o}}$, 其尺寸与输入特征尺寸一致.

2.4　基于 TLF 的时空特征聚类

同一视频的表观特征和运动特征经过 FIM 后, 每种模态的特征都从另一种模态的特征中获得了补充信息. 然而, 由于提取来源不同, 融合得到的表观和运动特征彼此之间必然存在差异; 此外, 虽然它们属于不同模态的特征, 但在本质上代表同一视频的相关信息. 因此, 本文希望网络可以忽略 2 种模态特征的差异, 并从中学习到一致信息. 对于相同模态的特征, 相同类别的视频必然具有相似的特征. 例如, 若视频 A 和视频 B 包含的行为类别一致, 那么 2 个视频最终提取到的表观或运动特征也应该相似. 本文设计了特征共享模块, 并添加基于不同模态和基于相同模态特征的 2 个 TLF 对网络进行约束, 实现时空特征的聚类.

2.4.1　特征共享模块

尽管表观特征和运动特征来源不同, 但两者属于相互增强的关系, 根据 2 类特征进行分类时应得到相同的识别结果. 本文设计一种特征共享模块, 在训练过程中共享参数, 从 2 种模态特征中提取一致性信息, 其结构如图 9 所示. 可以看出, FIM 的输出特征 $\boldsymbol{f}^{\mathrm{r}}$ 和 $\boldsymbol{f}^{\mathrm{o}}$ 经特征共享模块后得到特征 $\boldsymbol{R}$ 和 $\boldsymbol{O}$.

2.4.2　时空特征聚类

为了促进表观特征和运动特征之间的协同合作, 本文提出的 STN 采用双流分支结构同时从 RGB 帧和光流帧上提取表观和运动特征, 并通过

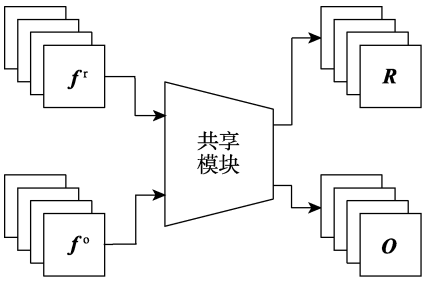


图 9　特征共享模块示意图

特征共享模块从 2 种模态特征中提取一致信息, 使同一视频的 2 种模态特征 $\boldsymbol{R}$ 和特征 $\boldsymbol{O}$ 在高维空间上接近. 为了使不同标签的 2 个视频中的一个视频的表观特征 $\boldsymbol{R}$ 与另外一个视频的运动特征 $\boldsymbol{O}$ 在高维空间上相距较远, 本文设计了 2 种样本选择策略来构造 TLF 训练网络. 网络训练时, 其输入包括 2 个三元组, 其中一个三元组包括 RGB 流视频帧、光流帧正样本和光流帧负样本, 另一个三元组则包括光流帧、RGB 流视频帧正样本和 RGB 流视频帧负样本.

基于第 1 个三元组输入, 本文设计了不同模态特征之间的 TLF, 其由 2 部分构成, 表示为

$$\begin{aligned}\mathrm{TLF}_1 = &\sum_{i=1}^{B}\left[\left\|\boldsymbol{R}_i^a-\boldsymbol{O}_i^p\right\|_2^2-\left\|\boldsymbol{R}_i^a-\boldsymbol{O}_i^n\right\|_2^2+\alpha_1\right]_+ + \\ &\sum_{i=1}^{B}\left[\left\|\boldsymbol{O}_i^a-\boldsymbol{R}_i^p\right\|_2^2-\left\|\boldsymbol{O}_i^a-\boldsymbol{R}_i^n\right\|_2^2+\alpha_2\right]_+ \end{aligned} \tag{3}$$

其中, $\boldsymbol{R}$ 和 $\boldsymbol{O}$ 分别表示经过特征共享模块后得到的表观和运动特征, 上标 a 表示用作参考的特征(锚点), p 表示与锚点特征来自同一标签视频的另一模态特征(正样本), n 表示与锚点特征来自不同标签视频的另一模态特征(负样本); α_1 和 α_2 均表示阈值. 通过该损失函数可以将拥有同一标签、不同视频的不同模态特征在高维空间上的距离拉近, 使共享模块从 2 种模态的特征中学习一致性信息. TLF 的作用机理如图 10 所示.

图 10 中, 正方形表示从 RGB 流中提取的表观特征, 三角形表示从光流帧中提取的运动特征. 式(3)中, 第 1 项以 RGB 流表观特征作为锚点, 减小光流运动特征正样本与锚点在高维特征空间中的距离, 增大运动特征负样本与锚点的距离; 第 2 项以光流运动特征为锚点, 减小 RGB 流表观特征正样本与锚点的距离, 增大 RGB 流特征负样本与锚点的距离. 通过这种形式的 TLF 可以将相同类别标签视频的表观特征与运动特征聚类到一起, 促使网络从 2 种模态特征中提取一致信息. 时空特征的聚类结果如图 11 所示.

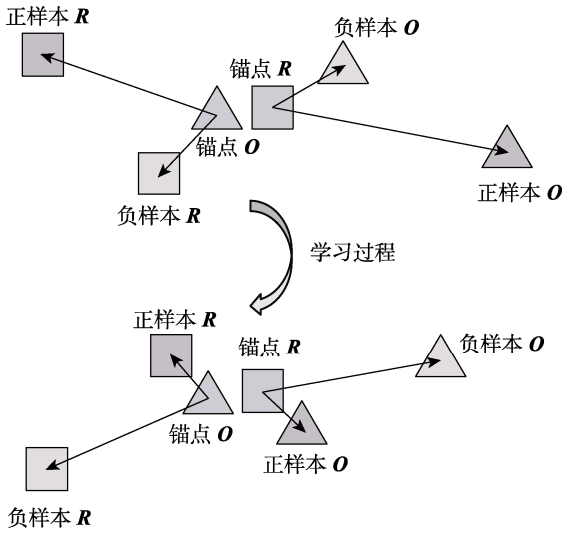


图 10 时空特征聚类过程示意图

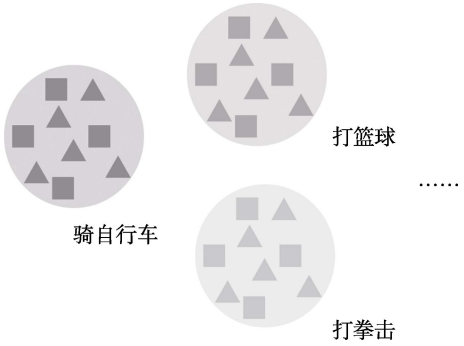


图 11 基于式(3)的时空特征聚类结果示意图

TLF_1 可以使网络学到 2 种模态特征的潜在关系，但 2 种模态的信息又各自具有优势. 如表观特征能够捕捉到空间的视觉信息，运动特征可以捕捉运动物体的速度、朝向等与时间维度相关的信息. 本文希望2种模态特征可以在一定程度上保持自身特有的信息，并共同作用，提高识别的准确度. 本文还设计了一种同模态特征内部的 TLF，公式为

$$\mathrm{TLF}_2=\sum_{i=1}^{B}\left[\left\|\boldsymbol{R}_i^a-\boldsymbol{R}_i^p\right\|_2^2-\left\|\boldsymbol{R}_i^a-\boldsymbol{R}_i^n\right\|_2^2+\alpha_3\right]_+ + \sum_{i=1}^{B}\left[\left\|\boldsymbol{O}_i^a-\boldsymbol{O}_i^p\right\|_2^2-\left\|\boldsymbol{O}_i^a-\boldsymbol{O}_i^n\right\|_2^2+\alpha_4\right]_+ \tag{4}$$

与式(3)不同，式(4)中每一项仅考虑同一模态特征之间的距离，目的是减小拥有相同标签、不同视频下的同模态特征间的距离，增大不同标签视频、同模态特征之间的距离，使不同模态特征保持其特有的信息. 对每一种模态的特征进行分类，将这种特有的信息反映在分类结果上. 最后融合基于 2 种模态特征的分类得分，进一步加强模型的判别能力. 基于式(4)的时空特征聚类结果如图 12 所示.

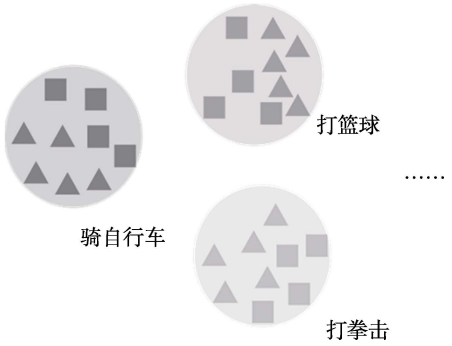


图 12 基于式(4)的时空特征聚类结果示意图

2.4.3 难样本挖掘三元组损失

实现 TLF 的关键在于如何得到三元组，这个过程称为三元组挖掘. 三元组挖掘最早采用离线的方法，即在网络训练前，通过统计数据标签的方式决定输入的三元组样本数据，然而此方法由于需要遍历所有数据，并且需要定期离线地更新三元组，因此其效率较低. 另一种方式采用在线的三元组挖掘方式，即根据输入的一个批量里所有的样本标签来选择三元组. 如在包含 b 个样本的批量数据中可以得到 b^3 个三元组，根据损失函数公式来计算 TLF 训练网络. 这种方式效率更高，因此本文采用在线挖掘的方式选择三元组样本.

TLF 公式表示为

$$\mathrm{TLF}=\mathrm{Max}\left(D(a,p)-D(a,n)+\alpha,\ 0\right).$$

其中，$D(x,y)$ 表示样本 x 和 y 之间的距离；a 表示锚点特征；p 表示正样本特征；n 表示负样本特征；α 表示距离阈值. 其核心思想是使负样本与锚点的距离大于正样本与锚点的距离. 三元组一般可以划分为简单三元组、半困难三元组和困难三元组 3 种类型，如图 13 所示. 其中，简单三元组是指锚点与正样本的距离小于锚点与负样本的距离且损失小于等于 0 的三元组；半困难三元组指锚点与正样本的距离大于锚点与负样本的距离且损失大于 0 的三元组；困难三元组指负样本比正样本更加靠近锚点的三元组，此类三元组往往是决定网络性能的关键. 因此，本文采用难样本挖掘的策略来选择困难三元组样本输入网络，通过计算与锚点最近的负样本与锚点之间的距离 $D(a,n)$，以及与锚点最远的正样本与锚点之间的距离 $D(a,p)$ 得到损失函数值，反向传播更新网络参数.

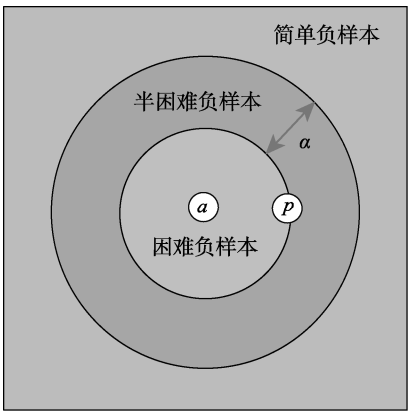


图 13　难样本挖掘示意图

3　实验及结果分析

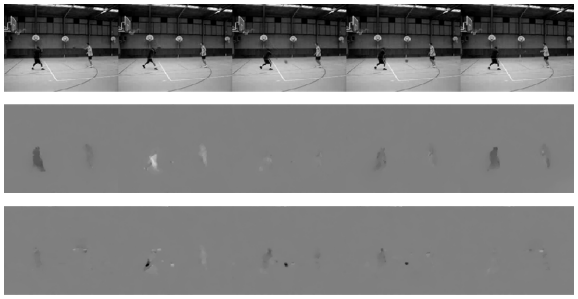
为了验证本文方法的有效性, 在 2 个广泛使用的公开数据集 UCF101[25]和 HMDB51[26]上进行实验, 并与现有的行为识别方法进行对比.

3.1　数据集及评测指标

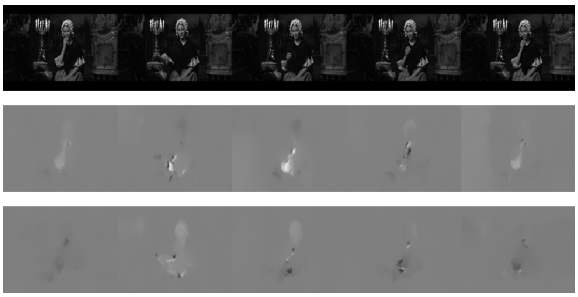
UCF101 数据集由中佛罗里达大学构建, 视频取自 Youtube 的剪辑视频. 该数据集包含 101 个现实的人类行为种类, 共计 13 320 个视频片段; 其中的行为可分组为身体运动、运动、人与物体交互、人与人交互和乐器演奏 5 类; 在动作方面提供了较大的多样性, 并且在场景、相机视点、物体外观姿势和规模, 以及照明条件等方面具有较大差异. 因此, UCF101 被认为是具有挑战性的数据集之一.

HMDB51 数据集由布朗大学的 Serre 实验室构建, 视频数据来源于 YouTube 和电影剪辑视频, 包含 51 个行为种类, 每个类别至少有 100 段具有合理差异的视频数据, 共计 6 766 个视频片段; 其中的动作可以分为面部动作(咀嚼、大笑、说话、微笑), 身体动作(拍手、走路、挥手、跳跃、拉起、侧手翻等), 面部与物体交互动作(抽烟、进食、喝水), 身体动作(梳头发、挥剑、倒水、接球、骑自行车等)和人际互动(击剑、踢腿、握手等)5 类. HMDB51 数据集中视频的场景和行为复杂多变, 极具挑战性.

图 14 所示为来自 2 个数据集的部分视频帧数据以及提取的光流帧数据. 图 14a 中, 第 1 行代表 RGB 视频帧, 第 2 行和第 3 行分别代表 x 方向和 y 方向上的光流帧; 图 14b 中, 第 5 行和第 6 行代表 x 方向和 y 方向上的光流帧. 从图 14 中可以看出, 虽然 RGB 图像表观相差较大, 但两者的光流特征存在一定的相似性, 说明了利用 2 种模态特征的互补信息进行行为识别的必要性.



a. UCF101 数据集 RGB 图像及其光流图像



b. HMDB51 数据集 RGB 图像及其光流图像

图 14　2 个数据集中的部分数据

本文采用行为识别中最常用的评测指标——Top1 准确度和 Top5 准确度来评价算法性能. 其中, Top1 准确度表示若模型对当前样本预测的最高概率的类别与样本真实标签相符, 则判定模型预测正确; Top5 准确度表示若模型对当前样本预测概率前 5 名的类别里包含有标签类别, 则认为模型判断正确. 准确度计算公式为 $A=T/(T+F)$. 其中, T 表示所有测试样本中被正确预测的个数; F 表示所有测试样本中被错误预测的个数.

3.2　实验结果与分析

3.2.1　MMSNet 实验

图像分类任务与行为识别任务存在一定的差异, 为了验证 MMSNet 在行为识别任务中的有效性, 本文未在大规模公开图像数据集 ImageNet 上进行实验, 而是在行为识别数据集 UCF101 和 HMDB51 上进行训练和测试, 并与现有的特征提取网络进行对比.

TSN[4]是 Two-stream 网络[14]的改进版本, 解决了行为识别中的长时依赖性问题. 由于双流结构的 TSN 模型性能优异, 且其 2 个分支分别采用 RGB 图像和光流图像作为输入, 除特征提取网络部分外没有其他冗余结构, 能够很好地评价特征提取网络的性能. 因此, 本文选用 TSN 模型作为

行为识别的基线模型进行实验, 实验结果如表 1 和表 2 所示. 从表 1 和表 2 可以看出, 与 ResNet[19], DenseNet[20]和 MobileNetV2[27]等经典的特征提取网络相比, MMSNet 参数量较少, 准确度最高. 特别地, MMSNet 与专为移动端图像分类设计的轻量级特征提取网络 MobileNetV2[27]参数量基本一致, 但准确度更高. 如当仅使用 RGB 图像输入时, MMSNet 在 UCF101 数据集上的 Top1 准确度为 87.45%, 比 MobileNetV2 提高了 25.77 个百分点.

表 1 不同方法在 UCF101 数据集实验结果对比

方法	$10^{-6}\times$ 参数量	RGB/%		光流/%		双流/%	
		Top1	Top5	Top1	Top5	Top1	Top5
VGG19[18]	21.3	41.25	76.07	41.26	76.15	41.26	76.18
ResNet101[19]	42.6	55.91	81.19	57.95	79.30	71.57	88.78
DenseNet101[20]	18.2	66.75	86.93	59.53	77.30	75.93	91.67
MobileNetV2[27]	2.3	61.68	85.93	57.66	68.91	68.91	90.60
MMSNet	2.2	87.45	96.63	71.60	91.00	92.48	98.04

表 2 不同方法在 HMDB51 数据集实验结果对比

方法	$10^{-6}\times$ 参数量	RGB/%		光流/%		双流/%	
		Top1	Top5	Top1	Top5	Top1	Top5
VGG19[18]	21.3	17.99	39.55	17.99	29.55	17.99	29.40
ResNet101[19]	42.6	21.66	46.66	18.36	30.49	23.76	58.91
DenseNet101[20]	18.2	29.36	58.21	22.06	42.63	32.12	71.34
MobileNetV2[27]	2.3	19.33	57.41	17.27	30.71	26.74	68.28
MMSNet	2.2	40.19	66.72	24.53	50.55	42.59	78.24

在未使用预训练模型的情况下, 当同时使用 RGB 和光流图像作为输入时(双流输入), MMSNet 的准确度均超过其他特征提取网络. 在 UCF101 和 HMDB51 数据集上的 Top1 准确度分别达到了 92.48%和 42.59%, 优于用于图像分类任务的特征提取网络(如 ResNet[19], DenseNet[20]等), 证明了 MMSNet 的有效性.

为验证所设计的轻量级卷积结构的有效性, 本文使用不同的卷积结构构造 MMSNet, 并在 UCF101 数据集上进行实验, 结果如表 3 所示. 可以看出, 与标准卷积相比, 本文设计的卷积结构在使用 RGB 图像作为输入时, Top1 和 Top5 准确度分别提升了 2.34 个百分点和 0.78 个百分点. 在采用双流输入并融合 2 个分支的分类得分后, 本文方法比标准卷积在 Top1 准确度指标上提升了 1.85 个百分点, 比深度可分离卷积提升了 4.44 个百分点, 进一步证明本文设计的轻量级卷积结构的有效性.

表 3 在 UCF101 数据集不同卷积结构实验结果对比

方法	$10^{-9}\times$ 计算量	RGB/%		光流/%		双流/%	
		Top1	Top5	Top1	Top5	Top1	Top5
标准卷积	0.46	85.11	95.85	66.98	87.23	90.63	98.04
Howard 等[22]	0.05	80.16	98.69	64.75	88.30	88.04	98.11
MMSNet	0.32	87.45	96.63	71.60	91.00	92.48	98.04

为了验证所设计的卷积结构的效率, 本文计算了在输入尺寸为(1,256,32,32)的特征图时不同卷积结构的计算量. 从表 3 可以看出, 与第 2.2.2 节的论述相同, 本文设计的卷积结构计算量高于 Howard 等[22]提出的深度可分离卷积, 但低于标准卷积, 且准确度最高. 也就是说, 本文提出的卷积结构在计算量和精度之间达到了较优的权衡.

3.2.2 基于特征交互和聚类的行为识别方法实验

为了验证本文方法的有效性, 首先在 UCF101 数据集上开展消融实验, 并与现有的行为识别方法进行对比, 实验结果如表 4 所示. 表 4 中的数据均为 Top1 准确度. 可以看出, 采用 MMSNet 进行特征提取的 TSN, 与采用 ResNet101 的 TSN 模型相比, 在双流输入的情况下, Top1 准确度提升了 20.91 个百分点. 添加 FIM 构建 STN 后, 准确度提升至 93.61%. 进一步地, 利用本文所设计的 TLF 训练 STN 后, 准确度提升了 1.21 个百分点. 更重要的是, 本文方法在未使用预训练模型的情况下, 与 TSN(ResNet101)相比, 在 UCF101 数据集上的准确度提升了 23.25 个百分点, 这进一步证明了本文

表 4 UCF101 数据集上 15 种方法 TOP1 准确度对比 %

行为识别方法	RGB	光流	双流
Two-stream[14]	51.26	48.17	51.86
C3D[9]	41.36	44.91	59.94
C3D(ResNet101)[9]	51.13	42.61	61.57
R(2+1)D[10]	62.27	51.26	62.49
TSN(ResNet101)[4]	55.91	57.98	71.57
TSF(VGG16)[15]	82.61	86.25	90.62
LTC[28]			92.70
ST-ResNet[16]			93.40
iDT + VLMPF[29]			94.30
HFV-ST-ResNet[30]			94.30
ResNet3D[31]			94.50
LVAR[11]			93.80
TSN(MMSNet)	87.45	71.60	92.48
STN(MMSNet)			93.61
STN + TLF			94.82

方法的有效性. 另外, 与基于3D卷积的行为识别方法 C3D(ResNet101)和 R(2+1)D 相比, 本文方法在 UCF101 数据集上的准确度分别提高了 33.25 和 32.33 个百分点.

为了验证本文方法的泛化性, 在 HMDB51 数据集上进行实验, 结果如表 5 所示. 可以看出, 在 TSN 结构下, 将特征提取网络 ResNet101 更换为 MMSNet, 在双流输入时 Top1 准确度提升了 18.83 个百分点; 利用 TLF 训练 STN 后, 其 Top1 准确度比 TSN 提升了 20.27 个百分点, 比改进的 TSN(MMSNet)提高了 1.44 个百分点. 这些证明了本文方法具有泛化性.

表 5　HMDB51 数据集上的 8 种方法 TOP1 准确度对比　%

行为识别方法	RGB	光流	双流
Two-stream[14]	17.99	11.87	17.99
C3D[9]	21.34	13.01	36.59
C3D(ResNet101)[9]	26.55	13.70	32.89
R(2+1)D[10]	36.56	17.99	37.82
TSN(ResNet101)[4]	21.66	18.36	23.76
TSN(MMSNet)	40.19	24.53	42.59
STN(MMSNet)			43.31
STN + TLF			44.03

4　结　语

针对现有的行为识别方法难以捕获时空特征互补信息的问题, 本文提出一种基于特征交互和聚类的 STN 以进行准确的行为识别. 首先采用 2 个混合多尺度分支, 分别从 RGB 流和光流输入数据中提取时空特征; 然后采用基于 Non-local 操作的 FIM 加强时空特征的交互, 增强特征的判别性; 最后基于相同视频的不同模态特征在分类时应该是相互增强的理念, 设计了特征共享模块和基于难样本挖掘的 TLF, 使网络能从 2 种模态特征中学习一致信息, 实现对时空特征的聚类. 进一步, TLF 通过拉近同标签视频不同模态特征的距离, 使共享模块从 2 种模型信息中提取行为的共同表示; 通过拉近相同模态特征在高维空间的距离, 使不同模态特征保持其独立的一些信息, 在分类时互相补充以达到更加准确的分类结果. 实验结果表明, 本文方法有效地提高了基线方法的识别准确度, 在 UCF101 数据集上 Top1 准确度达到 94.82%, 在 HMDB51 数据集上达到 44.03%, 能够高效、准确地进行行为识别.

参考文献(References):

[1] Herath S, Harandi M, Porikli F. Going deeper into action recognition: a survey[J]. Image and Vision Computing, 2017, 60: 4-21

[2] Wang X L, Girshick R, Gupta A, *et al*. Non-local neural networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7794-7803

[3] Schroff F, Kalenichenko D, Philbin J. Facenet: a unified embedding for face recognition and clustering[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 815-823

[4] Wang L M, Xiong Y J, Wang Z, *et al*. Temporal segment networks: towards good practices for deep action recognition[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 20-36

[5] Laptev I, Marszalek M, Schmid C, *et al*. Learning realistic human actions from movies[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2008: 1-8

[6] Wang H, Schmid C. Action recognition with improved trajectories[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2013: 3551-3558

[7] Donahue J, Hendricks L A, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 2625-2634

[8] Ji S W, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231

[9] 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. Los Alamitos: IEEE Computer Society Press, 2015: 4489-4497

[10] Tran D, Wang H, Torresani L, *et al*. A closer look at spatiotemporal convolutions for action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6450-6459

[11] Chang Y L, Chan C S, Remagnino P. Action recognition on continuous video[J]. Neural Computing and Applications, 2021, 33(4): 1233-1243

[12] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018

[13] Shi L, Zhang Y, Cheng J, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 12026-12035

[14] 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. Heidelberg: Springer, 2014: 568-576

[15] 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. Los Alamitos: IEEE Computer Society Press, 2016: 1933-1941

[16] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal multiplier networks for video action recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 7445-7454

[17] Cho S, Foroosh H. Spatio-temporal fusion networks for action recognition[C] //Proceedings of the Asian Conference on Computer Vision. Heidelberg: Springer, 2018: 347-364

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2021-11-10]. https://arxiv.org/abs/1409.1556

[19] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[20] Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2261-2269

[21] Chang X B, Hospedales T M, Xiang T. Multi-level factorisation net for person re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2109-2118

[22] Howard A G, Zhu M, Chen B, *et al*. Mobilenets: efficient convolutional neural networks for mobile vision applications[OL]. [2021-11-10]. https://arxiv.org/abs/1704.04861

[23] Buades A, Coll B, Morel J M. A non-local algorithm for image denoising[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2005: 60-65

[24] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010

[25] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild[OL]. [2021-11-10]. https://arxiv.org/abs/1212.0402

[26] Kuehne H, Jhuang H, Garrote E, *et al*. HMDB: a large video database for human motion recognition[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 2556-2563

[27] Sandler M, Howard A, Zhu M, *et al*. Mobilenetv2: inverted residuals and linear bottlenecks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 4510-4520

[28] Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1510-1517

[29] Duta I C, Ionescu B, Aizawa K, *et al*. Spatio-temporal vector of locally max pooled features for action recognition in videos[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3205-3214

[30] Butt A M, Yousaf M H, Murtaza F, *et al*. Agglomerative clustering and residual-VLAD encoding for human action recognition[J]. Applied Sciences, 2020, 10(12): 4412

[31] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6546-6555