第 10 卷 第 5 期
2021 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 10 No. 5
Sep. 2021

引文格式：
夏鼎, 王亚立, 乔宇. 基于人体部件的视频行为识别方法研究 [J]. 集成技术, 2021, 10(5): 23-33.
Xia D, Wang YL, Qiao Y. Research of video action recognition based on human body parts [J]. Journal of Integration Technology, 2021, 10(5): 23-33.

基于人体部件的视频行为识别方法研究

夏 鼎[1,2] 王亚立[1] 乔 宇[1*]

[1]（中国科学院深圳先进技术研究院 深圳 518055）
[2]（中国科学院大学 北京 100049）

摘 要 现有人体行为识别算法主要依赖于粗粒度的视频特征，但这些特征不足以有效描述人体行为的动作构成，从而降低了深度学习模型对易混淆行为的识别能力。该研究提出了一种基于人体部件的视频行为识别方法，通过学习人体细粒度部件的动作表示，自底向上地学习人体行为视频表征。该方法主要包含：(1) 部件特征增强模块，用于增强基于图像的人体部件特征；(2) 部件特征融合模块，用于融合人体各部件特征以形成人体特征；(3) 人体特征增强模块，用于增强视频帧中所有人的人体特征。该方法在国际标准数据库 UCF101 和 HMDB51 上进行的实验验证结果显示，基于人体部件的视频行为识别方法与已有方法具有良好的互补性，可以有效提高人体行为识别精度。

关键词 行为识别；深度学习；人体部件；特征融合
中图分类号 TP 399 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20210308001

Research of Video Action Recognition Based on Human Body Parts

XIA Ding[1,2] WANG Yali[1] QIAO Yu[1*]

[1](*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[*]*Corresponding Author*: *yu.qiao@siat.ac.cn*

Abstract High-level action information, such as spatial feature of frames, temporal feature among frames, or human level skeleton features are usually used in existing video action recognition methods. However, these high-level features cannot effectively describe the action composition of human behavior, and thus reduce the ability of deep learning models to recognize confusing behaviors. In this work, video action recognition method based on human body parts is investigated. By learning the action representation of the fine-grained parts of the human body, the video representation of human action was learned from bottom to up level. Specifically, the method mainly includes three modules: (1) body part feature enhancement module,

收稿日期：2021-03-08 **修回日期**：2021-04-14
作者简介：夏鼎，硕士研究生，研究方向为视频行为分析；王亚立，副研究员，研究方向为计算机视觉；乔宇（通讯作者），研究员，研究方向为计算机视觉，E-mail：yu.qiao@siat.ac.cn。

which enhances the image-based human body part feature, (2) body part feature fusion module, which fuses the features of various parts of the human body to form human feature, and (3) body feature enhancement module, which is responsible for enhancing the human body features of all people in the video. The popular datasets of UCF101 and HMDB51 were used for experiments. And the results showed that, the video action recognition method based on human body parts is complementary with current methods, and can effectively improve the accuracy of human action recognition.

Keywords action recognition; deep learning; body part; feature fusion

1 引 言

行为识别是视频任务的一个基础问题，在越来越多的领域中得到应用，如智能监控、机器人交互、视频推荐等。近年来，随着深度学习的发展以及大规模视频数据库的出现，行为识别研究逐步走向成熟。

目前，大多数的方法都是将行为识别看作视频分类，关注于更好地利用视频帧的静态特征与帧之间的动态特征。Karpathy 等[1]提出使用一个二维卷积神经网络，独立地提取每一帧的特征，然后融合时间信息的方法。然而，该方法没有考虑帧之间的动态特征。基于此，Simonyan 和 Zisserman[2]提出了双流架构——一个二维卷积神经网络学习静态图片特征，另一个二维卷积神经网络学习光流动态特征。此方法的双流融合策略比较简单，并被广泛扩展到其他方法。其中，时序分段网络（Temporal Segement Network，TSN）[3]采用稀疏采样策略来学习长视频片段特征，并利用光流探索了不同组合策略来融合两个流的特征。但计算光流对计算资源消耗太大，所以时空运动网络（SpatioTemporal and Motion，STM）[4]采用一个模块来学习帧级别的动态特征，节省了计算光流的消耗。与 STM 直接将时空特征与动态特征相加不同，时序激活与聚合网络（Temporal Excitation and Aggregation，TEA）[5]利用动态特征来重校准时空特征，从而加强学习得到的运动模式。除了二维卷积神经网络，三维神经网络也被设计用于学习动态特征。C3D[6]模型将 2D 图像模型扩展到时空间领域，用相似的方法处理空间与时间维度。然而，3D 模型参数较多，对计算资源要求较高。R（2＋1）D[7]模型将 3D 卷积分解成 2D 空间卷积加上 1D 时间卷积，节省计算资源。在 3D 模型的基础上，为了更好地利用时间特征，SlowFast[8]使用 Slow 流与 Fast 流捕捉不同的空间与时间特征。这些方法都关注于视频帧级别的高层次特征，易忽视掉动作的细节特征，于是有些方法开始利用更细粒度的特征来提升视频行为分类效果。例如，Wang 和 Gupta[9]利用 Region Proposal Network[10]检测出人体位置后，使用图卷积神经网络提取人体级别特征，进行视频行为识别。之后，活动者关系图网络（Actor Relation Graph，ARG）[11]也探索了利用图卷积神经网络提取人体特征的群体行为识别方法。

由于人体是由人体部件组成，且一个人的行为是由每个部件的动作组合而成，所以本研究提出一种利用人体部件特征进行视频行为识别的方法。该方法更深入地挖掘出行为相关的细粒度特征并加以利用，与其他行为识别方法提取的高层特征（如视频帧特征）具有很好的互补性，可以有效提升视频行为识别精度。最终，在 UCF101[12]与 HMDB51[13]两个标准数据集上验证了人体部

件特征对于视频行为识别任务的有效性。本文方法具有以下优点：(1)提出了提取人体部件特征的方法用于视频行为识别；(2)利用更细粒度的人体部件特征，将人体行为视为人体部件动作在时空上的组合，有效提升传统视频行为识别方法的精度。

2　基于人体部件的视频行为识别方法

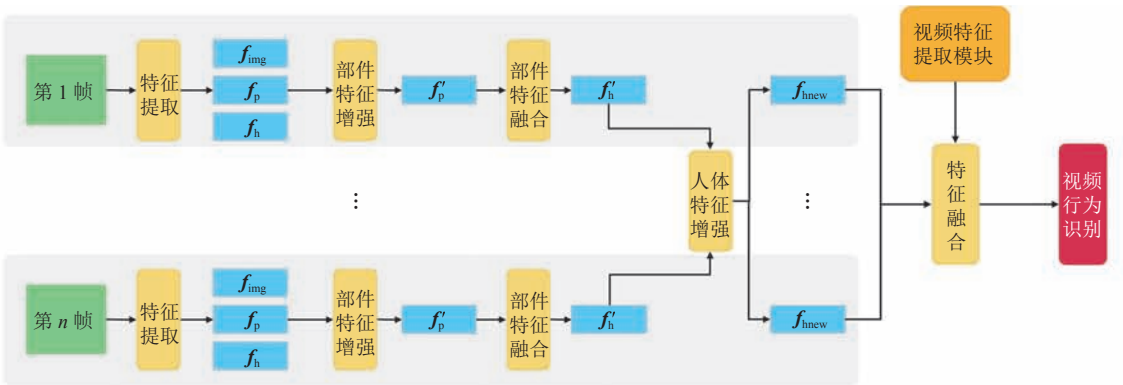
视频行为识别的目的是识别出视频中人类的行为类别，相对于静态图像中的行为识别任务而言，视频行为识别任务的复杂度更高。该任务的主要难点是，视频中人类行为具有类间与类内方差。具体来讲，人类可以在不同的视角下以不同的速度做出同样的行为。另外，一些具有相似动作模式的行为难以区分。传统上，视频行为识别任务被认作是一个高层视频分类任务。目前，主流的二维卷积神经网络方法仍关注于提取更有效的视频帧之间时空特征。然而，这种方法忽视了对人类行为细节的理解。人类行为是由一系列人体部件动作在时空上的组合形成，通过自底向上地理解人体部件动作以及它们的时空组合方式，可以更好地理解人类的高层语义行为。本研究通过提取人体部件特征，自底向上地将人体部件动作组合成人体行为，利用更细粒度的行为特征提升视频行为识别效果。模型架构如图 1 所示。

2.1　特征提取

目前，视频行为数据集没有针对人体部件位置的标注，所以本研究通过自动化的方法计算出每个人体部件的位置后，从视频帧提取特征，具体流程如图 2 所示。首先，使用在 Microsoft COCO[14]数据集上预训练的 KeypointRCNN[15]模型检测出每个人的位置框与关节点，之后通过关节点计算出每个人体部件的位置框。使用到的人体部件有：头、左手臂、左手掌、右手臂、右手掌、臀、左腿、左脚、右腿、右脚。然后，使用在 ImageNet[16]数据集上预训练的 ResNet50[17]模型提取视频帧特征，将倒数第 2 个全连接层特征作为视频帧特征，记为 $\boldsymbol{f}_{\text{img}}$。同时，在 stage1 输出的特征图上，根据人体位置框与人体部件位置框，利用 RoIAlign[10]技术提取出人体特征与人体部件特征，维度为 256×7×7，之后通过两层卷积神经网络，重排列成 2 048 维特征向量，记为 $\boldsymbol{f}_{\text{h}}$、$\boldsymbol{f}_{\text{p}}$。

2.2　网络模块

该网络由 3 个模块组成：部件特征增强模块、部件特征融合模块和人体特征增强模块。其



f_{img}：图像特征；f_{p}：人体部件特征；f_{h}：人体特征；f'_{p}：增强的人体部件特征；f'_{h}：融合形成的人体特征；f_{hnew}：增强的人体特征

图 1　模型架构

Fig. 1　Model framework

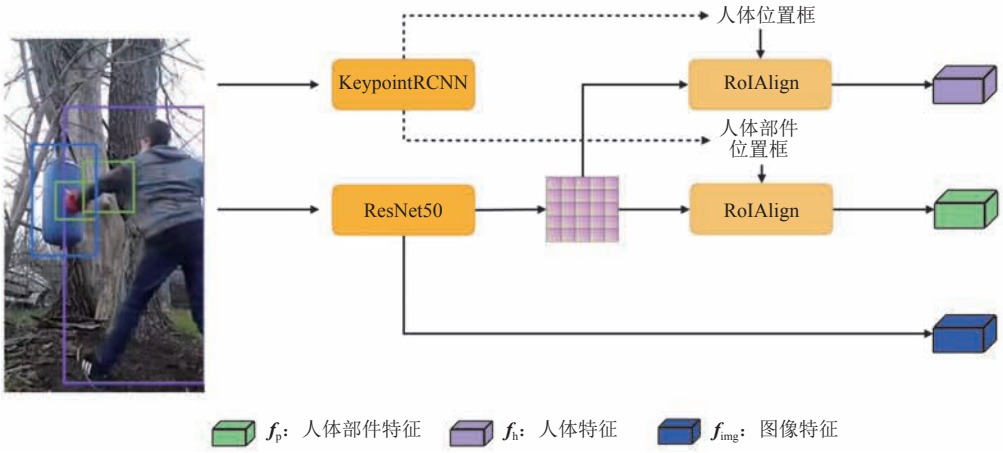


图 2 特征提取流程

Fig. 2 Pipeline of feature extracting

中，部件特征增强模块负责利用人体特征与视频帧特征等粗粒度的环境信息增强人体部件特征；部件特征融合模块负责将人体部件特征融合成人体特征；人体特征增强模块利用图卷积神经网络，学习视频中所有人之间的关系，从而获得视频级人体表征，用于行为识别。

2.2.1 部件特征增强模块

本文研究以人体部件特征为基础，所以需要增强直接从视频帧上提取的原始人体部件特征 $\boldsymbol{f}_{\mathrm{p}}$，作为整个算法的基础。相对人体部件而言，视频帧与人体都属于粗粒度环境信息。该模块通过拼接的方式，将视频帧特征与人体特征等环境信息融合到人体部件特征中，并经过两层全连接层以及一层 dropout[18]层进行降维处理，得到增强的人体部件特征 $\boldsymbol{f}'_{\mathrm{p}}$。全连接层中间使用 ReLU[19]作为激活函数。公式如下：

$$\boldsymbol{f}_1=concat\left(\boldsymbol{f}_{\mathrm{img}},\boldsymbol{f}_{\mathrm{h}},\boldsymbol{f}_{\mathrm{p}}\right) \tag{1}$$

$$\boldsymbol{f}_2=\boldsymbol{f}_{c_2}\left\{relu\left[\boldsymbol{f}_{c_1}\left(\boldsymbol{f}_1\right)\right]\right\} \tag{2}$$

$$\boldsymbol{f}'_{\mathrm{p}}=dropout\left(\boldsymbol{f}_2\right) \tag{3}$$

其中，$concat(\cdot)$为特征拼接操作；$\boldsymbol{f}_{\mathrm{img}}$ 为视频帧特征；$\boldsymbol{f}_{\mathrm{h}}$ 为人体特征；$\boldsymbol{f}_{\mathrm{p}}$ 为人体部件特征；$\boldsymbol{f}'_{\mathrm{p}}$ 为增强的人体部件特征；$relu(\cdot)$为 ReLU[19]激活函数；$\boldsymbol{f}_1$、$\boldsymbol{f}_2$ 为中间层特征；$dropout$ 的概率取 0.3。第一层全连接层 $\boldsymbol{f}_{c_1}$ 将特征降维至 2 048 维，第二层全连接层 $\boldsymbol{f}_{c_2}$ 将特征降维至 512 维。

2.2.2 部件特征融合模块

部件特征融合模块的目标是将增强的各人体部件特征融合成人体特征。由于人体具有自然的分级结构，所以本模块自底向上地分级融合人体部件特征，过程如图 3 所示。首先，融合手掌和手臂的特征，得到手的特征，之后融合左右手与头的特征，得到上半身特征；然后，融合脚与大小腿的特征，得到腿的特征，随后融合左右腿与臀部的特征，得到下半身特征；最后，融合上半

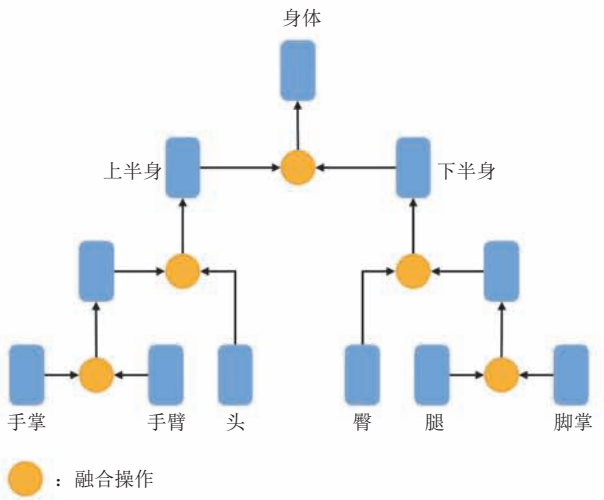


图 3 部件特征融合模块

Fig. 3 Body part feature fusion module

身与下半身特征，得到新的人体特征 $\boldsymbol{f}'_{\mathrm{h}}$。其中，融合操作由特征拼接操作与一层全连接层组成。

2.2.3 人体特征增强模块

由于视频行为识别需要综合视频中所有人的行为进行判断，所以人体特征增强模块负责综合所有人的行为与位置信息，增强每个人的人体特征，架构如图 4 所示。参考 ARG 网络，该模块使用人体的图像特征以及位置信息来学习每个特征之间的关联性，构建关系矩阵，计算公式如下：

$$\boldsymbol{G}_{ij}=\frac{R_s\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)exp\left[R_a\left(\boldsymbol{f}'_{\mathrm{h}_i},\boldsymbol{f}'_{\mathrm{h}_j}\right)\right]}{\sum_{j=1}^{N}R_s\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)exp\left[R_a\left(\boldsymbol{f}'_{\mathrm{h}_i},\boldsymbol{f}'_{\mathrm{h}_j}\right)\right]} \tag{4}$$

其中，$\boldsymbol{G}_{ij}$ 为关系矩阵第 i 行 j 列的值；N 为所有视频帧上的人体总数；$\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 表示第 i 和 j 个人的位置框左上角与右下角坐标；$\boldsymbol{f}'_{\mathrm{h}_i}$、$\boldsymbol{f}'_{\mathrm{h}_j}$ 表示融合后的第 i 和 j 个人的人体特征；$R_a\left(\boldsymbol{f}'_{\mathrm{h}_i},\boldsymbol{f}'_{\mathrm{h}_j}\right)$为对两个人之间的图像特征关系进行建模；$R_s\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$为对两个人之间的位置信息关系进行建模。

$$R_a\left(\boldsymbol{f}'_{\mathrm{h}_i},\boldsymbol{f}'_{\mathrm{h}_j}\right)=\frac{\theta\left(\boldsymbol{f}'_{\mathrm{h}_i}\right)^{\mathrm{T}}\phi\left(\boldsymbol{f}'_{\mathrm{h}_j}\right)}{\sqrt{d_k}} \tag{5}$$

$$R_s(\boldsymbol{x}_i,\boldsymbol{x}_j)=\mathrm{II}\left[d\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\leqslant\mu\right] \tag{6}$$

其中，$\mathrm{II}(\cdot)$为指示函数；$d(\cdot)$为计算两个位置框中心坐标的欧几里德距离；$\theta(\cdot)$、$\phi(\cdot)$为线性映射函数；d_k 为线性映射函数输出的特征维度；μ 为超参数，表示距离的阈值。

由于多个人之间的关系比较复杂，所以在此模块对所有人构建了多个关系矩阵 $\boldsymbol{G}^1,\boldsymbol{G}^2,\cdots,\boldsymbol{G}^{N_g}$，其中 N_g 表示关系矩阵的个数，每个关系矩阵不共享参数。之后利用图卷积神经网络进行信息传递过程，计算公式如下：

$$\boldsymbol{Z}_{\mathrm{out}}=\sum_{i=1}^{N_g}\sigma\left(\boldsymbol{G}^i\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{W}^i\right) \tag{7}$$

$$\boldsymbol{Z}=\boldsymbol{Z}_{\mathrm{out}}+\boldsymbol{Z}_{\mathrm{in}} \tag{8}$$

其中，$\boldsymbol{Z}_{\mathrm{out}}$ 为中间层特征；$\boldsymbol{Z}_{\mathrm{in}}$ 为由人体特征 $\boldsymbol{f}'_{\mathrm{h}}$ 组成的特征矩阵；$\boldsymbol{W}$ 为可学习的权重矩阵；$\sigma(\cdot)$为激活函数，本文使用 ReLU[19]函数；$\boldsymbol{Z}$ 为增强的人体特征组成的特征矩阵，记增强的人体特征为 $\boldsymbol{f}_{\mathrm{newh}}$。流程如图 4 所示，每一层的计算如公式(7～8)所示。

3　训练网络

训练网络的主要功能是提取人体部件的细粒度行为特征，融合成人体特征 $\boldsymbol{f}_{\mathrm{newh}}$。在视频行为识别任务中，还需要结合作为环境信息的视频帧特征 $\boldsymbol{f}_{\mathrm{frame}}$。记提取视频帧特征的网络为视频特征提取网络，其可以使用 TSN、时序偏移网络

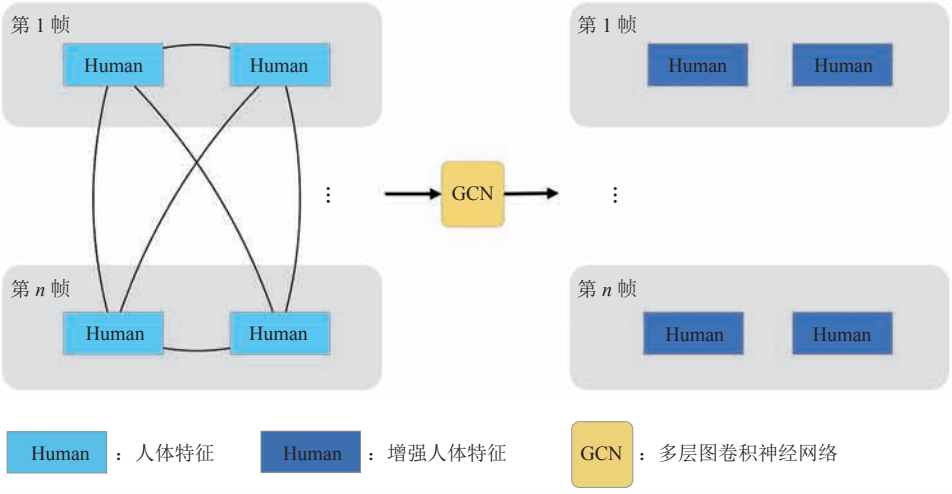


图 4　人体特征增强模块

Fig. 4　Body feature enhancement module

(Temporal Shift Module，TSM)[20]、STM、TEA等经典的二维卷积神经网络。由于数据集的大小、数据分布不同，训练网络可能造成过拟合问题，所以本文采用以下两种策略：在出现过拟合问题的数据集上采取独立训练策略，在未出现过拟合问题的数据集上采取联合训练策略。

(1)联合训练。将每一视频帧上的所有人体特征经过平均池化层后，与视频帧特征拼接，之后经过全连接层预测出每一帧的行为分数。随后将所有帧上的行为分数经过平均池化层后得到预测的视频行为分数，取分数最大的类别为预测的视频行为类别。此时，该网络与视频特征提取网络联合训练参数。计算公式如下：

$$\boldsymbol{f}_{\text{human}}=pooler_{\text{mean}}\left(\boldsymbol{f}_{\text{newh}}\right) \tag{9}$$

$$\boldsymbol{f}'_{\text{frame}}=concat\left(\boldsymbol{f}_{\text{frame}},\boldsymbol{f}_{\text{human}}\right) \tag{10}$$

$$\hat{\boldsymbol{y}}_v=pooler_{\text{mean}}\left[\boldsymbol{f}_c\left(\boldsymbol{f}'_{\text{frame}}\right)\right] \tag{11}$$

$$Loss=L_v\left(\boldsymbol{y}_v,\hat{\boldsymbol{y}}_v\right) \tag{12}$$

其中，$pooler_{\text{mean}}(\cdot)$为平均池化操作；$concat(\cdot)$为特征拼接操作；$\boldsymbol{f}'_{\text{frame}}$为中间层特征；$\hat{\boldsymbol{y}}_v$为视频行为的预测类别标签；$\boldsymbol{y}_v$表示视频行为的真实类别标签。损失函数$Loss$使用交叉熵函数$L_v$。

(2)独立训练。与联合训练类似，该策略不拼接$\boldsymbol{f}_{\text{frame}}$、$\boldsymbol{f}_{\text{human}}$，而是使用它们独立训练两个网络。测试时将两个网络预测的视频行为分数相加，作为最终视频行为分数的预测值。

4 实验与讨论

本研究使用两个经典的视频行为识别数据集HMDB51与UCF101来验证方法效果。接下来将介绍使用的数据库，以及针对各数据集的实验结果进行分析。

4.1 数据库介绍

HMDB51数据集是通过各种渠道搜集到的真实视频组成的视频集，如电影和网络视频，包含51个动作类别的6 766个视频片段。本文试验使用标准的split1划分方式划分训练与测试集。UCF101数据集包含101个动作类别，13 320个视频片段。本文实验在两个数据集都采取标准的split1划分方式划分训练与测试集。

4.2 对比实验

首先，本文在HMDB51数据集上进行实验。由于该数据集比较小，联合训练会造成严重的过拟合问题，所以在该数据集上的实验采取独立训练策略。基础网络分别使用TSN、TSM、STM、TEA网络作为对比基准，它们的主干网络都选择ResNet50[17]，并在ImageNet[16]数据集上进行预训练。训练时每个视频等间隔采样8帧，每帧选取面积最大的3个人。学习率为0.001，在第30和60个epoch时分别下降至0.1倍。采用带动量的随机梯度下降策略，momentum为0.9，同时加上L2正则化，参数为0.000 5。将两个网络分别训练好后，测试时将二者的预测值经过Softmax归一化后相加，数值最大的类别为预测的视频行为类别。表1结果显示，对于4种基础网络，加上本文方法后都能取得精度提升。例如，使用TSN时，本文方法精度提升了1.43%；使用TEA时，达到了过拟合状态，然而使用了本文方法后精度提升了2.61%。该实验结果表明，基于人体部件的细粒度行为特征与基于视频帧的粗粒度行为特征具有很好的互补性，且能部分解决过拟合问题。

其次，在UCF101数据集上进行实验。由于

表1 HMDB51数据集对比实验

Table 1 HMDB51 dataset ablation study

基础网络	精度（%）	
	WithoutOurs	WithOurs
TSN	52.29	53.72
TSM	52.35	53.79
STM	48.24	52.55
TEA	51.05	53.66

该数据集较大，不会造成过拟合，所以采取联合训练策略。基础网络同样采用 TSN、TSM、STM、TEA 作为对比基准，它们的主干网络都选择 ResNet50，并在 ImageNet 数据集上进行预训练。学习率为 0.01，在第 30、45 和 55 个 epoch 时分别下降至 0.1 倍。其他参数与在 HMDB51 上的实验相同。采用联合训练策略，将两个网络联合训练好后，测试时将视频行为预测值经过 Softmax 归一化，选取数值最大的类别为预测的视频行为类别。表 2 结果显示，对于 4 种基础网络，本文方法都能取得精度提升。例如，使用 TSN 时，本文方法精度提升了 2.67%；使用 TEA 时，本文方法精度提升了 1.22%。该实验结果同样表明基于人体部件的细粒度行为特征的有效性。

表 2　UCF101 数据集对比实验

Table 2　UCF101 dataset ablation study

基础网络	精度（%）	
	WithoutOurs	WithOurs
TSN	84.59	87.26
TSM	86.15	86.76
STM	83.00	83.24
TEA	82.92	84.14

4.3　结果可视化分析

首先，在 HMDB51 测试集上选取性能提升最大的 20 个动作类别进行详细的精度对比。采用的基础网络为 TSN，结果如图 5 所示。从图 5 可知，在与人体部件动作强相关的行为类别上，本文方法取得了较大的精度提升，如 shoot_ball、throw、jump、pushup 等。图 6 是基础网络为 TEA 时的对比结果，在 shake_hands、climb、shoot_ball 等与肢体动作强相关的类别上，本文方法同样取得了较大的精度提升。

然后，在 UCF101 数据集的实验中，选取基础网络为 TSN 时性能提升最大的 20 个动作类别的精度进行对比。图 7 结果显示，在与人体部件动作强相关的类别上，本文方法都能取得较大的分类精度提升，如 BodyWeightSquats、HandstandWalking。图 8 是基础网络为 TEA 时的对比结果，在 JumpingJack、Lunges 等与肢体动作强相关的类别上，本文方法同样取得了较大的精度提升。

由此可知，在与人体部件动作强相关的类别中，细粒度的人体部件特征能与粗粒度的视

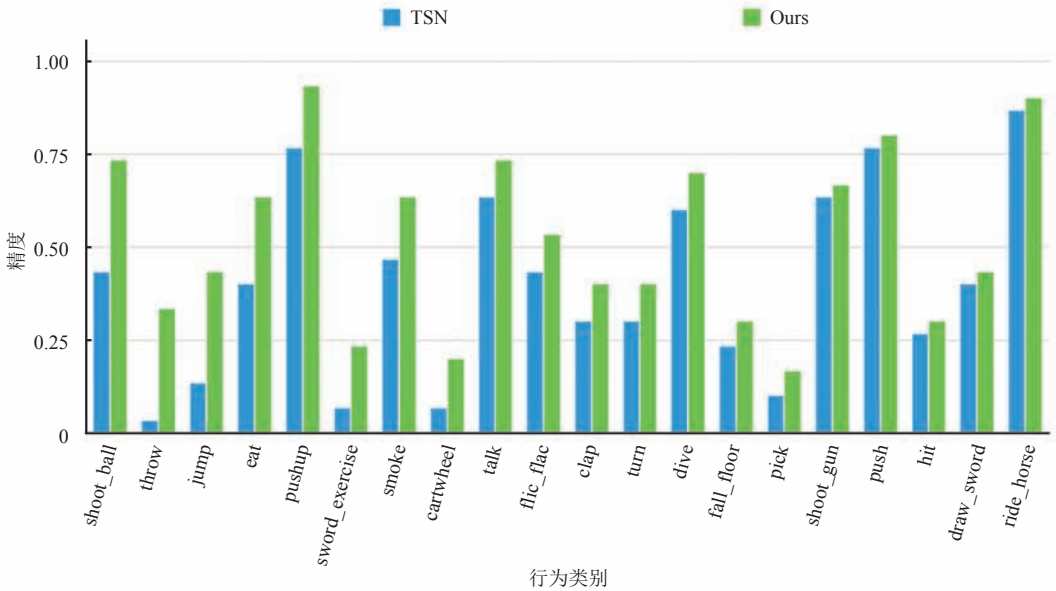


图 5　本文方法与 TSN 网络在 HMDB51 测试集上的类别精度对比

Fig. 5　HMDB51 class accuracy comparison of ours and TSN

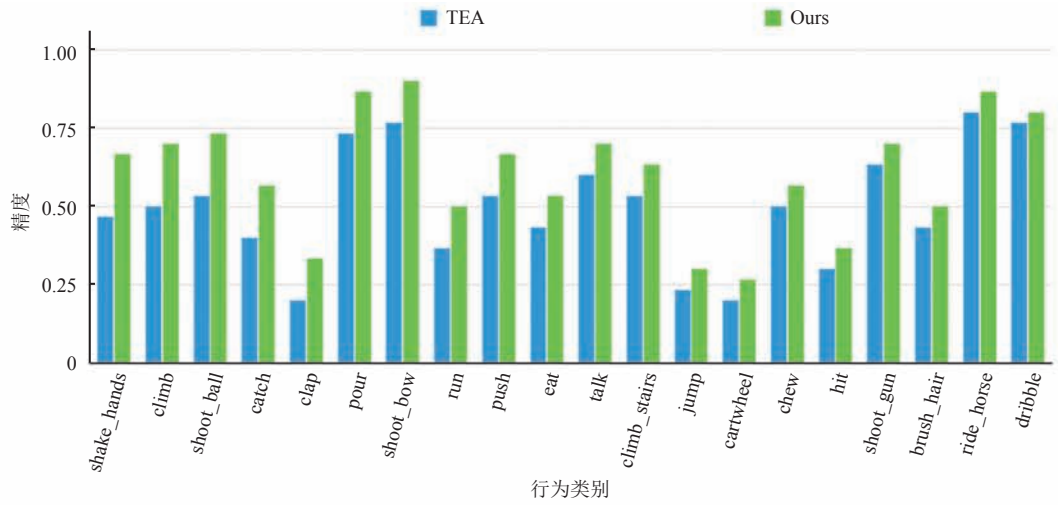


图 6　本文方法与 TEA 网络在 HMDB51 测试集上的类别精度对比

Fig. 6　HMDB51 class accuracy comparison of ours and TEA

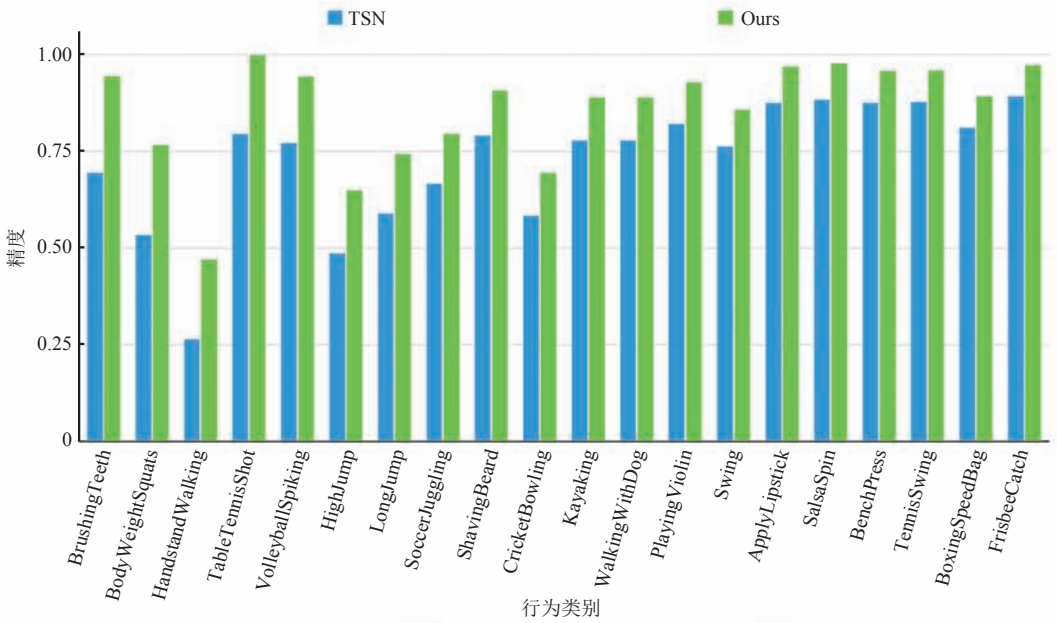


图 7　本文方法与 TSN 网络在 UCF101 测试集上的类别精度对比

Fig. 7　UCF101 class accuracy comparison of ours and TSN

频帧特征形成良好的互补，提升视频行为分类效果。

4.4　案例分析

相较于传统方法，本文方法在大部分行为类别上有比较大的精度提升，然而在部分类别上依然有所不足。在 UCF101 数据集上，本文方法在 Haircut、HammerThrow 等类别上的错误相对较多，如图 9 所示。其中，Haircut 失败样例大部分都是误识别为 BlowDryHair，从图 9 样例可知，Haircut 与 BlowDryHair 的主要区别是手持工具的不同，由于手掌级别的人体部件动作相似，当算法不能识别出工具时，本文方法就容易进行误识别。对于 HammerThrow 类别，失败样例大部分都是误识别为 ThrowDiscus。通过样例分析可知，该方法主要通过物体与场景进行两个类别的区分。当不能识别出物体时，主要依靠场景进行

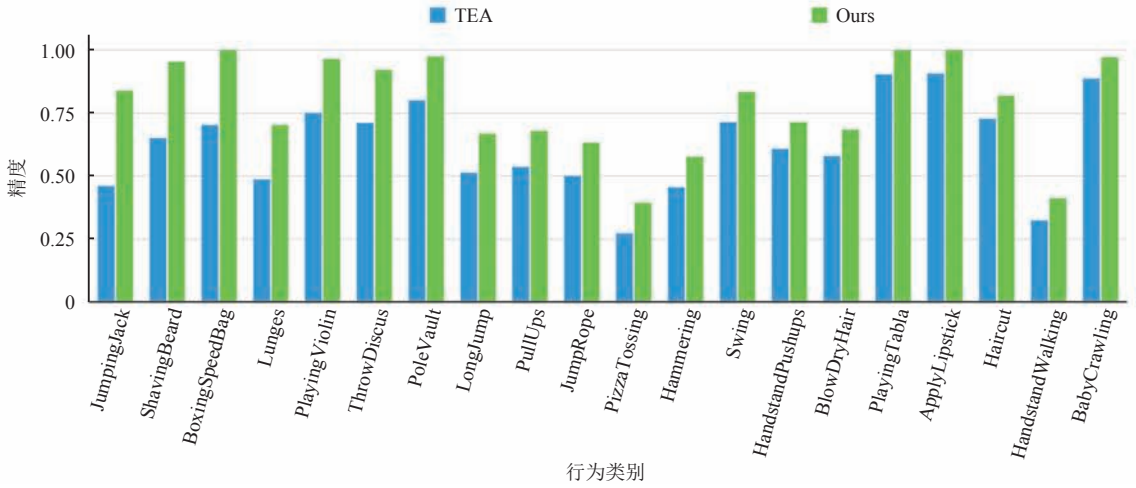


图 8　本文方法与 TEA 网络在 UCF101 测试集上的类别精度对比

Fig. 8　UCF101 class accuracy comparison of ours and TEA

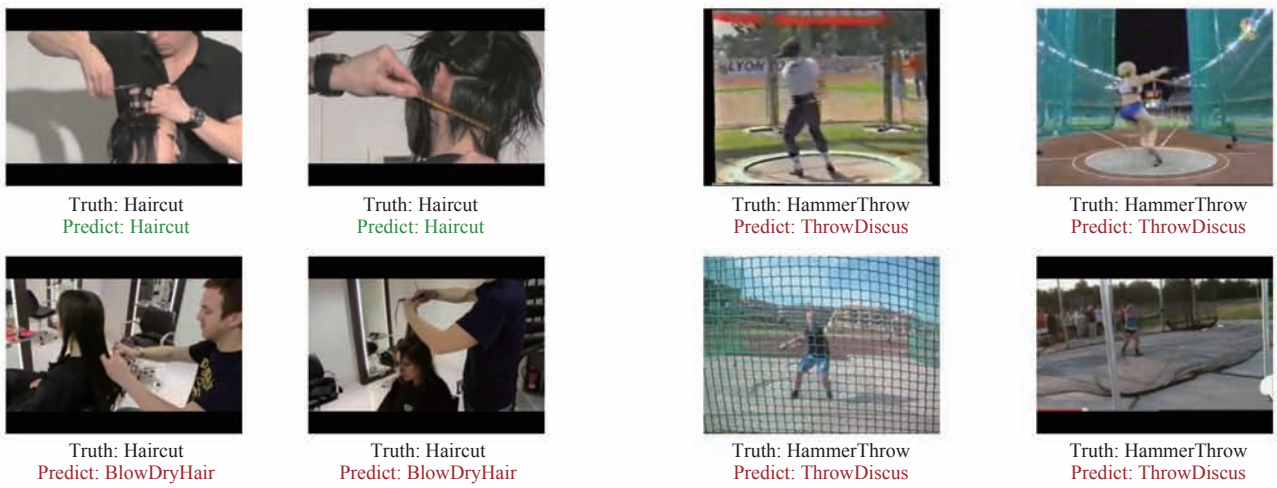


图 9　UCF101 样例分析

Fig. 9　UCF101 examples analysis

类别区分。在 HammerThrow 类别中，场景都比较相似，本文方法不能识别出该场景，故容易误识别为 ThrowDiscus。

从图 9 可知，在与人体部件动作强相关的类别上，本文方法精度有较好的提升，但在识别人体部件动作相似，主要依靠物体与场景区分的易混淆类别时，本文方法表现不足。由于本文使用的人体部件特征没有更精细化到手指等级别，所以在部分人体部件动作相似的类别上仍然需要依靠物体与场景进行识别。

5　讨论与分析

现阶段对视频行为识别的研究，主要集中在使用粗粒度特征的领域，如视频帧特征或人体特征。然而每个人体部件的细节动作都与视频行为相关，这些方法没有更进一步对细粒度的人体部件特征进行探索，易忽视掉许多行为细节。在不使用 Kinetics[21]数据集预训练，且只使用视频帧的图像数据条件下，TSN 方法在 UCF101 数据集和 HMDB51 数据集上的精度分别为 84.59%、

52.29%，此时只使用视频帧特征进行视频行为识别的方法均达到过拟合的状态。融合本文方法后，精度分别为 87.26%、53.72%，分别取得 2.67%、1.43% 的提升。可见，对于视频行为识别任务来说，人体部件特征与视频帧特征具有非常好的互补性，同时能够部分解决过拟合问题。本文初步提出一种使用人体部件特征的有效方法，然而相对于 ARG 等网络使用人体特征的方法，本文研究利用人体部件特征的方法显得比较直接。

6 结 论

针对现有视频行为识别使用特征层次过高，在关注动作细节的类别上效果不是很好的问题，本文提出一种基于人体部件特征的视频行为识别方法。该方法可以对人体部件进行自动定位，提取基于人体部件的细粒度行为特征，从而有效地进行视频行为识别。同时，本文介绍了两种训练策略，针对不同的数据集可以采取更合适的对应策略。实验结果表明，使用不同的基础网络，本文方法都可以有效地提升视频行为识别效果。在实际的视频行为数据中，每个人体部件都有各自的动作，然而当前的数据集的标注只针对视频级别或者帧级别进行标注，没有提供每个人体部件级别的监督信息，所以大部分的方法都没有细化到提取人体部件特征。本文初步证明了人体部件特征的有效性。未来工作中，将重点关注人体部件特征融合方式的探索，以及如何更有效地将细粒度的人体部件特征与粗粒度的视频帧特征结合起来，以提升视频行为识别效果。

参 考 文 献

[1] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1725-1732.

[2] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [Z/OL]. arXiv: 1406.2199, 2014.

[3] Wang LM, Xiong YJ, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2016: 20-36.

[4] Jiang BY, Wang MM, Gan WH, et al. STM: spatiotemporal and motion encoding for action recognition [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 2000-2009.

[5] Li Y, Ji B, Shi XT, et al. TEA: temporal excitation and aggregation for action recognition [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 909-918.

[6] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2015: 4489-4497.

[7] 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, 2018: 6450-6459.

[8] Feichtenhofer C, Fan HQ, Malik J, et al. SlowFast networks for video recognition [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6202-6211.

[9] Wang XL, Gupta A. Videos as space-time region graphs [C] // Proceedings of the European Conference on Computer Vision, 2018: 399-417.

[10] Ren SQ, He KM, Ross G, et al. Faster R-CNN: towards real-time object detection with region proposal networks [Z/OL]. arXiv: 1506.01497, 2015.

[11] Wu JC, Wang LM, Wang L, et al. Learning actor relation graphs for group activity recognition [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9964-9974.

[12] Soomro K, Zamir AR, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild [Z/OL]. arXiv: 1212.0402, 2012.

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

[14] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context [Z/OL]. arXiv: 1405.0312, 2014.

[15] He KM, Gkioxari G, Dollár P, et al. Mask R-CNN [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[16] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[17] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[18] Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [Z/OL]. arXiv: 1207.0580, 2012.

[19] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines [C] // International Conference on Machine Learning, 2010: 807-814.

[20] 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, 2019: 7083-7093.

[21] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset [Z/OL]. arXiv: 1705.06950, 2017.