JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
12
VOL.25

ISSN1006-8961
CN11-3758/TB

数字浮雕建模 *P2647*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第12期（总第296期）

2020年12月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

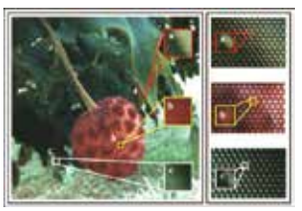
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

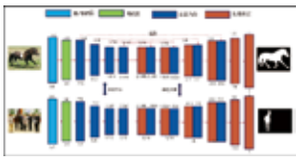
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年12月 第25卷 第12期（总第296期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

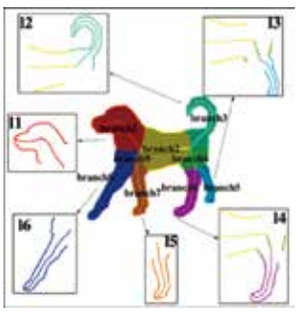
目次



光场显著性检测研究综述(第2465页)



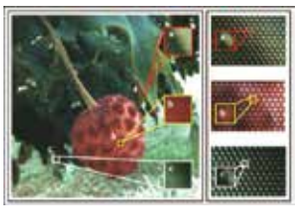
融合注意力机制与知识蒸馏的孪生网络压缩(第2563页)



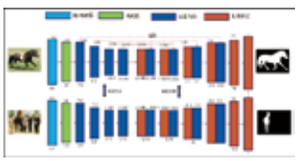
面向可解释性的物体拓扑结构骨架表征方法(第2587页)

Volume 25, Number 12
Published December 16, 2020

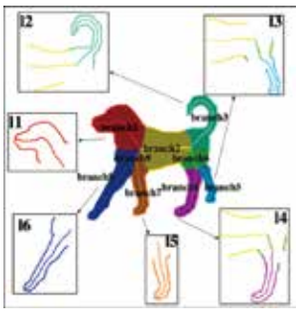
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Review of saliency detection on light fields(P2465)



Combining attention mechanism and knowledge distillation for Siamese network compression (P2563)



Skeleton characterization of object topology toward explainability (P2587)

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2020)12-2541-12

论文引用格式:Tan D T, Li S C, Chang W W and Li D L. 2020. Multi-feature fusion behavior recognition model. Journal of Image and Graphics, 25(12):2541-2552(谭等泰,李世超,常文文,李登楼. 2020. 多特征融合的行为识别模型. 中国图象图形学报,25(12):2541-2552)[DOI:10.11834/jig.190637]

多特征融合的行为识别模型

谭等泰[1,2],李世超[1],常文文[3],李登楼[2]

1. 甘肃政法大学公安技术学院,兰州　730070;2. 甘肃政法大学司法鉴定中心,兰州　730070;
3. 兰州交通大学电子与信息工程学院,兰州　730070

摘　要:**目的**　视频行为识别和理解是智能监控、人机交互和虚拟现实等诸多应用中的一项基础技术,由于视频时空结构的复杂性,以及视频内容的多样性,当前行为识别仍面临如何高效提取视频的时域表示、如何高效提取视频特征并在时间轴上建模的难点问题。针对这些难点,提出了一种多特征融合的行为识别模型。**方法**　首先,提取视频中高频信息和低频信息,采用本文提出的两帧融合算法和三帧融合算法压缩原始数据,保留原始视频绝大多数信息,增强原始数据集,更好地表达原始行为信息。其次,设计双路特征提取网络,一路将融合数据正向输入网络提取细节特征,另一路将融合数据逆向输入网络提取整体特征,接着将两路特征加权融合,每一路特征提取网络均使用通用视频描述符——3D ConvNets(3D convolutional neural networks)结构。然后,采用 BiConvLSTM(bidirectional convolutional long short-term memory network)网络对融合特征进一步提取局部信息并在时间轴上建模,解决视频序列中某些行为间隔相对较长的问题。最后,利用 Softmax 最大化似然函数分类行为动作。**结果**　为了验证本文算法的有效性,在公开的行为识别数据集 UCF101 和 HMDB51 上,采用 5 折交叉验证的方式进行整体测试与分析,然后针对每类行为动作进行比较统计。结果表明,本文算法在两个验证集上的平均准确率分别为 96.47% 和 80.03%。**结论**　通过与目前主流行为识别模型比较,本文提出的多特征模型获得了最高的识别精度,具有通用、紧凑、简单和高效的特点。

关键词:行为识别;双路特征提取网络;3 维卷积神经网络;双向卷积长短期记忆网络;加权融合;高频特征;低频特征

Multi-feature fusion behavior recognition model

Tan Dengtai[1,2], Li Shichao[1], Chang Wenwen[3], Li Denglou[2]

1. School of Public Security and Technology, Gansu University of Political Science and Law, Lanzhou 730070, China;
2. GSIPSL Center of Judicial Expertise, Gansu University of Political Science and Law, Lanzhou 730070, China;
3. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

Abstract: **Objective** With the rapid development of internet technology and the increasing popularity of video shooting equipment (e.g., digital cameras and smart phones), online video services have shown an explosive growth. Short videos have become indispensable sources of information for people in their daily production and life. Therefore, identifying how these people understand these videos is critical. Videos contain rich amounts of hidden information as these media can store

收稿日期:2019-12-07;**修回日期**:2020-04-03;**预印本日期**:2020-04-10

基金项目:国家自然科学基金项目(61861002);甘肃省科技厅青年科学基金项目(17JR5RA159,18JR3RA192);甘肃省教育厅项目(2019B-119);甘肃政法大学重点项目(GZF2018XZDLW17);甘肃政法大学司法鉴定中心科研资助项目(jdzxyb2018-06)

Supported by: National Natural Science Foundation of China (61861002)

more information compared with traditional ones, such as images and texts. Videos also show complexity in their space-time structure, content, temporal relevance, and event integrity. Given such complexities, behavior recognition research is presently facing challenges in extracting the time domain representation and features of videos. To address these difficulties, this study proposes a behavior recognition model based on multi-feature fusion. **Method** The proposed model is mainly composed of three parts, namely, the time domain fusion, two-way feature extraction, and feature modeling modules. The two- and three-frame fusion algorithms are initially adopted to compress the original data by extracting high- and low-frequency information from videos. This approach not only retains most information contained in these videos but also enhances the original dataset to facilitate the expression of original behavior information. Second, based on the design of a two-way feature extraction network, detailed features are extracted from videos through the positive input of the fused data to the network, whereas overall features are extracted through the reserve input of these data. A weighted fusion of these features is then achieved by using the common video descriptor, 3D ConvNets (3D convolutional neural networks) structure. Afterward, BiConvLSTM (bidirectional convolutional long short-term memory network) is used to further extract the local information of the fused features and to establish a model on the time axis to address the relatively long behavior intervals in some video sequences. Softmax is then applied to maximize the likelihood function and to classify the behavioral actions. **Result** To verify its effectiveness, the proposed algorithm was tested and analyzed on public datasets UCF101 and HMDB51. Results of a five-fold cross-validation show that this algorithm has average accuracies of 96.47% and 80.03% for these datasets, respectively. Comparative statistics for each type of behavior show that the classification accuracy of the proposed algorithm is approximately equal in almost all categories. **Conclusion** Compared with the available mainstream behavior recognition models, the proposed multi-feature model achieves higher recognition accuracy and is more universal, compact, simple, and efficient. The accuracy of this model is mainly improved via two- and three-frame fusions in the time domain to facilitate video information analysis and behavior information expression. The network is extracted by a two-way feature to efficiently determine the spatio-temporal features of videos. The BiConvLSTM network is then applied to further extract the features and establish a timing relationship.

Key words: behavior recognition; two-way feature extraction network; 3D convolutional neural networks (3D ConvNets); bidirectional convolutional long short-term memory network (BiConvLSTM); weighted fusion; high-frequency feature; low-frequency feature

0 引 言

传统视频监控中,行为目标分析主要依赖人工审核,需要耗费大量人力和财力,且准确的分析非常耗时。因此对视频中的目标行为进行自动分析与识别具有重要的理论研究价值和广阔的应用前景,可以有效净化互联网环境,促进网络视频的监管(马钰锡 等,2019)。

行为识别的目的是通过计算机准确地自动分析未知视频中正在进行的特定行为。按照特征提取方式的不同,行为识别方法分为基于人工选取特征的方法和基于深度学习的方法。在基于人工选取特征的方法中,通常通过手动设计某些特定的特征,对数据集中特定的行为进行识别,典型的方法有梯度直方图(histograms of oriented gradients, HOG)(Harris 和 Stephens,1988)、光流直方图(histograms of optical flow, HOF)(Lowe, 1999)、时空兴趣点(space-time interest points, STIP)(Laptev 和 Lindeberg, 2003; Willems 等, 2008)、运动边界直方图(motion boundary histograms, MBH)(Scovanner 等,2007)及其在 3 维方向上的扩展,如 HOG3D(Kläser 等, 2008)、3D-SIFT(Yang 等,2019)等。Wang 等人将 MBH 成功用于密集轨迹算法(dense trajectories, DT)(Wang 等, 2011)和改进的密集轨迹算法(improving dense trajectories, iDT)(Wang 和 Schmid, 2013),iDT 算法采用密集采样策略提取特征点,提取密集轨迹点的 HOG 特征、HOF 特征、MBH 等局部特征。iDT 是基于人工选取特征的方法中准确率最高的算法。深度学习框架提出后,行为识别的准确率大幅提高,目前基于深度学习的主流结构有双流卷积网络(two stream),3 维卷积神经网络(3D convolutional neural networks, 3D ConvNets)和长短期记忆网络(long short-term memory, LSTM),如图 1 所示。

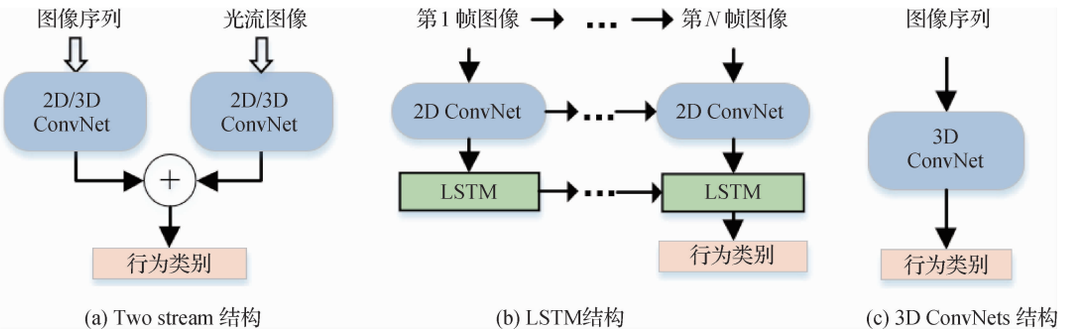


图 1　主流的行为识别结构

Fig. 1　Mainstream behavior recognition structure ((a) Two stream; (b) LSTM; (c) 3D ConvNets)

Two stream 结构(Simonyan 和 Zisserman,2014)主要对视频数据的稠密光流与 RGB 数据分别训练卷积神经网络(convolutional neural networks,CNN)模型,再将双流网络得到的结果进行融合。Ji 等人(2018)提出了一种端到端的结构,利用上下文信息、多任务学习等,在单个统一框架中直接输出语义细分,在动作识别方面,采用具有时间聚集的双流网络。为了解决双流网络中视频输入需要固定大小和固定长度的缺点,Wang 等人(2018)提出了一种端到端的双流融合网络,可以识别视频中任意尺度和长度的行为动作。针对大多数现有方法在空间和时间上都采用相同的网络结构,Chen 等人(2019)提出了时空异构双流网络,采用两种不同的网络结构获取时空信息。Two stream 结构可以有效应用于视频动作识别,尤其是在训练数据有限的情况下,能够很好地提取特征。

3D ConvNets(Tran 等,2015)结构是从传统的 2D ConvNets 网络拓展而来,在原来空间中加入时间维度,提取视频中的时空特征。为了在更深层次提取特征,研究人员对卷积核进行了改进。Diba 等人(2017)提出了 T3D 网络,使用基于 3D DenseNet 的动作识别架构和新的时间层以模拟可变时间卷积内核深度,但是这种方法仅在几个 RGB 帧上求值,忽略了时间信息在视频分析中的重要作用。Qiu 等人(2017)提出了 P3D ResNet 网络,用 1 × 3 × 3 卷积和 3 × 1 × 1 卷积代替 3 × 3 × 3 卷积,从而大幅减少计算量,与传统的 3D 卷积网络不同,此体系结构成功提高了视频识别任务的性能。LSTM 结构(Donahue 等,2015)利用卷积神经网络提取特征,采用 LSTM 在时间方向上建模,将空间相关性信息保持在 LSTM 过程中可以学习更多信息性的时空特征,进而进行行为识别。Ng 等人(2015)主要将双流法与 LSTM 进行结合,LSTM 网络进行时域信息建模,提升了模型对时域信息的表达能力。Sharma 等人(2015)将注意力机制的思想引入到行为识别中,思路简单,但是具有一定的启发意义。Liu 等人(2016)提出了一种基于树结构的遍历方法,在 LSTM 中引入了新的门控机制,以了解顺序输入数据的可靠性,并相应地调整其对更新存储单元中存储的长期上下文信息的影响。

此外,还有其他类型的方法。Zhu 等人(2016)通过提取视频中的关键帧提高了行为识别的准确率。Ouyang 等人(2019)提出了一种新颖的多任务学习架构,该架构结合 3D ConvNets 网络和 LSTM 网络以及多任务学习机制提取特征。Wang 等人(2019)采用传统的特征提取方法,使用交互式 3 维(interactive three dimensions,I3D)模型或其他一些模型的输出与改进的密集轨迹(improved dense trajectory,IDT)结合在一起,并通过词袋(bag of words,BoW)和 Fisher 向量(Fisher vector,FV)进行编码。Li 等人(2019)利用两流网络和时空层,获得了三通道融合模型,同时将轨迹约束应用于深层特征和手工特征,结合它们的优点,完成了视频事件识别。

尽管行为识别的研究取得了重要进展,但是由于视频时空结构的复杂性、内容的多样性以及时序的关联性,视频行为识别仍然存在很多需要解决的问题。本文提出了一种多特征融合的行为识别模型,通过时域融合模块压缩原始视频信息,能够完整表达行为信息,然后通过双路特征提取模块高效提取视频特征,最后通过双向卷积长短期记忆网络(bidirectional convolutional long short-term memory network, BiConvLSTM),解决某些行为动作间隔相

对较长的问题。

本文的贡献有:1)从时域信息融合的角度出发,提出了时域信息融合算法,将原始视频相邻帧中的高频信息和低频信息加权融合,将长视频综合成短视频,提高对视频信息的分析能力,使其能够完整表达行为信息,同时也增强了数据集。2)针对视频时空信息的特征提取,设计了双路特征提取网络,一路提取行为动作的细节特征,采用顺序视频作为网络的输入;另一路提取行为动作的整体特征,将逆序视频作为网络的输入。每一路特征提取网络均使用通用视频描述符——3D ConvNets 结构。3)为了解决视频时间序列中某些行为的执行方式不同、间隔相对较长等因素对行为识别的影响,设计了 BiConvLSTM 网络,对融合特征进一步提取局部细节特征并建立时序关系。

1 模型设计

本文提出的多特征融合的行为识别模型主要由时域融合模块、双路特征提取模块和特征建模模块 3 部分构成,模型的总体结构如图 2 所示。时域融合模块通过两帧融合算法和三帧融合算法压缩原始视频的数据量,减少原始视频的冗余信息,保留了原始视频绝大多数信息量。双路特征提取模块设计了两路 3D ConvNets 网络 Net0 和 Net1,Net0 将融合数据正向输入网络提取特征,Net1 将融合数据逆向输入网络提取特征,然后将两路特征加权融合。特征建模模块采用 BiConvLSTM 网络对融合特征进一步提取局部特征并建立时序关系。最后,利用 Softmax 最大化似然函数分类行为动作。

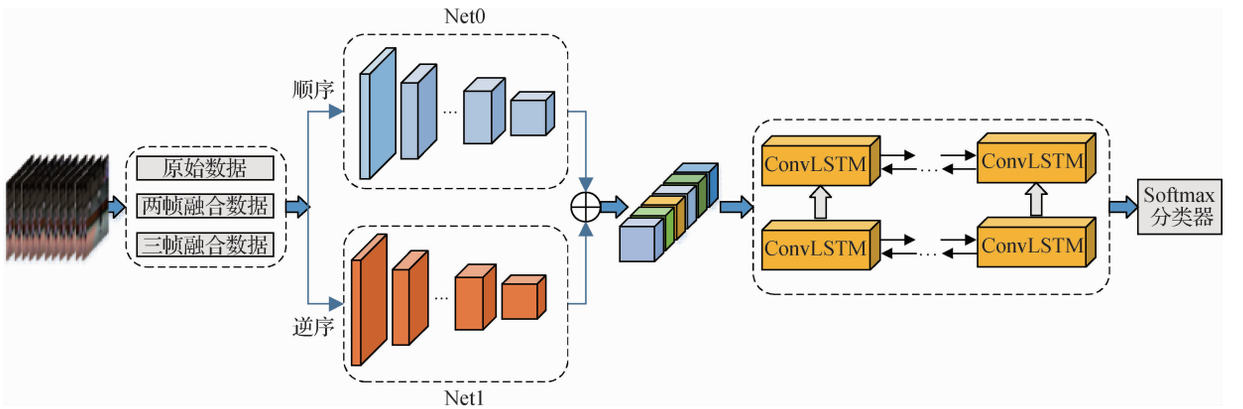


图 2 行为识别模型

Fig. 2 Behavior recognition model

2 时域融合模块

不仅信号存在高、低频分量,图像也存在高、低频分量。对于图像而言,低频分量表现的是整体结构,高频分量表现的是细节特征,显然,低频分量是存在冗余的。视频的高频分量是指视频帧不经过高斯滤波的原始视频;低频分量是指经过高斯滤波得到的视频。

利用高斯滤波提取视频中的低频信息,是通过正态分布计算图像中每个像素的变换,在 2 维空间的定义为

$$\boldsymbol{G}(u,v,\delta) = \frac{1}{2\pi\delta^2}\mathrm{e}^{-\frac{u^2+v^2}{2\delta^2}} \tag{1}$$

式中,δ 为尺度参数,δ 越大平滑越剧烈。假设 2 维图像为 $\boldsymbol{f}_k(x,y)$,则低频图像 $\boldsymbol{L}(x,y)$ 为二者卷积,即

$$\boldsymbol{L}(x,y) = \boldsymbol{G}(u,v,\delta) * \boldsymbol{f}_k(x,y) \tag{2}$$

时域融合模块的主要目的是降低原始视频的数据量,降低视频中含有的冗余信息,压缩原始视频的数据量。行为动作由一系列图像构成,含有大量与行为无关的信息。将图像序列记做 $\boldsymbol{f}_k(x,y)$,$k \in \mathbf{N}$,$\boldsymbol{f}_{k-1}(x,y)$ 和 $\boldsymbol{f}_{k+1}(x,y)$ 分别表示视频序列中的前一帧和后一帧图像。首先对当前帧 $\boldsymbol{f}_k(x,y)$ 使用高斯滤波器平滑获取图像的低频信息,然后将低频信息和高频信息 $\boldsymbol{f}_{k+1}(x,y)$ 重建得到融合图,其中,高频分量为其相邻帧,是不经过高斯滤波的原始图像。最后,在空间维度上进行下采样,减少每一帧图像的大小。

两帧融合过程的计算为

$$\boldsymbol{F}(x,y) = \lambda_1 \boldsymbol{L}(x,y) + \lambda_2 \boldsymbol{f}_{k+1}(x,y) \quad (3)$$

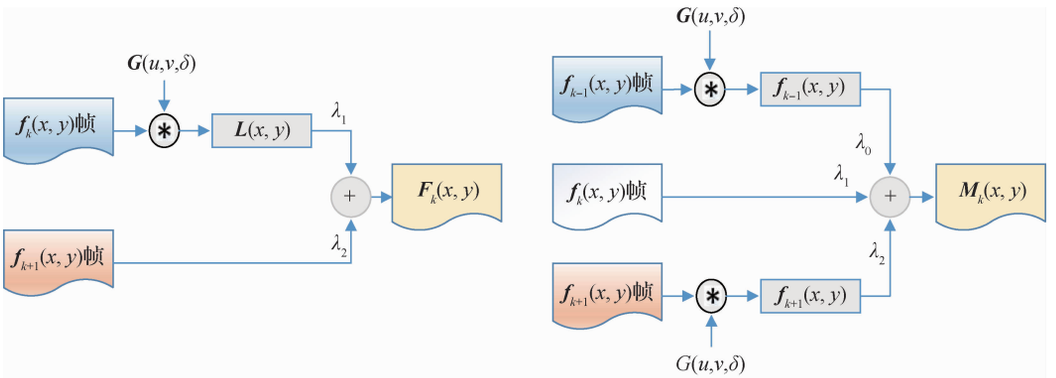
式中，$\boldsymbol{L}(x,y)$ 为图像的低频像素值，$\boldsymbol{f}_{k+1}(x,y)$ 为图像高频分量，$\boldsymbol{F}(x,y)$ 为融合后图像的像素值，λ_1 和 λ_2 为图像 $\boldsymbol{L}(x,y)$ 和 $\boldsymbol{f}_{k+1}(x,y)$ 像素的权重。本文 λ_1 和 λ_2 的取值分别为 0.6 和 0.4。

三帧融合过程的计算为

$$\boldsymbol{M}(x,y) = \lambda_0 \boldsymbol{L}_{k-1}(x,y) + \lambda_1 \boldsymbol{f}_k(x,y) + \lambda_2 \boldsymbol{L}_{k+1}(x,y) \quad (4)$$

式中，$\boldsymbol{f}_k(x,y)$ 为图像的低频像素值，$\boldsymbol{L}_{k-1}(x,y)$、$\boldsymbol{L}_{k+1}(x,y)$ 为图像高频分量，$\boldsymbol{M}(x,y)$ 为融合后图像的像素值，λ_0、λ_1 和 λ_2 为其权重，取值分别为 0.25、0.5 和 0.25。

两帧视频和三帧视频的融合过程如图 3 所示。



(a) 两帧融合过程　　(b) 三帧融合过程

图 3　视频融合

Fig. 3　Video fusion ((a) two frame fusion process; (b) three frame fusion process)

3　双路特征提取模块

3.1　3D ConvNets 原理

2D ConvNets 对图像具有很强的特征表达能力，但处理视频任务时，视频中相邻帧间具有很强的关联性，如果时间维度被转换为互不相关的图像帧，则容易丢失目标的运动信息。为了解决这一问题，本文采用了 3D ConvNets 结构，该结构能够很好地提取视频中与目标、场景、动作有关的特征。3D ConvNets 网络与 2D ConvNets 网络类似，由卷积层和下采样层交替堆叠而成，输入是多个连续帧构成的立方体，在立体空间尺度上提取特征，通过 3 维卷积核提取特征，从而捕获一段时间内的运动信息。3 维卷积核的计算为

$$v_{km}^{xyz} = f\left(b_{km} + \sum_{p=0}^{P_k-1}\sum_{q=0}^{Q_k-1}\sum_{r=0}^{R_k-1} w_{kmn}^{pqr} u_{(k-1)n}^{(x+p)(y+q)(z+r)}\right) \quad (5)$$

式中，v_{km}^{xyz} 为 3D ConvNets 的输出，u 为第 $k-1$ 层到第 k 层的输入，第 k 层卷积核的尺度为 $P_k \times Q_k \times R_k$，$f(\cdot)$ 为激活函数，b_{km} 为偏置，n 为索引，w_{kmn}^{pqr} 为权重系数。

与 2D ConvNets 网络一样，3D ConvNets 网络在时间和空间维度上分别进行 3 维下采样，减少卷积层之间的连接，逐步缩小特征图规模。常用的下采样方法有平均池化、最大池化等。最大池化是一种比较好的解决方案，具体为

$$v_{x,y,z} = \max_{0\leqslant i\leqslant s_1, 0\leqslant j\leqslant s_2, 0\leqslant k\leqslant s_3} (\boldsymbol{u}_{x\times s+i, y\times t+j, z\times r+k}) \quad (6)$$

式中，$\boldsymbol{u}$ 为 3 维输入向量，v 为池化层的输出，s、t、r 为步长。利用该方法采样后，特征图的尺寸在时间和空间维度都将缩小，计算量也随之减少，网络则变得更加鲁棒。

3.2　网络结构

本文使用的双路特征提取网络结构结合 Tran 等人(2015)网络的优点设计，如图 4 所示。

从图 4 可以看出，Net0 和 Net1 网络均由 6 个卷积层、4 个池化层、6 个激活函数(ReLU)和 3 个批标准化(batch normalization，BN)层组成。每个 Conv3D 层的卷积核大小为 3 × 3 × 3，6 个 Conv3D 层的滤波器个数依次为 32、64、128、128、256 和 256。除 Conv3D_3a 和 Conv3D_4a 外，每个 Conv3D 层后面都有 1 个 ReLU 图层和 1 个 BN 层。第 1 个池化层的内核大小为 2×2×1，步长为 2×2×1，在第 1 层 Conv3D 上只执行空间汇聚，保留时间维度信息量。其他池化

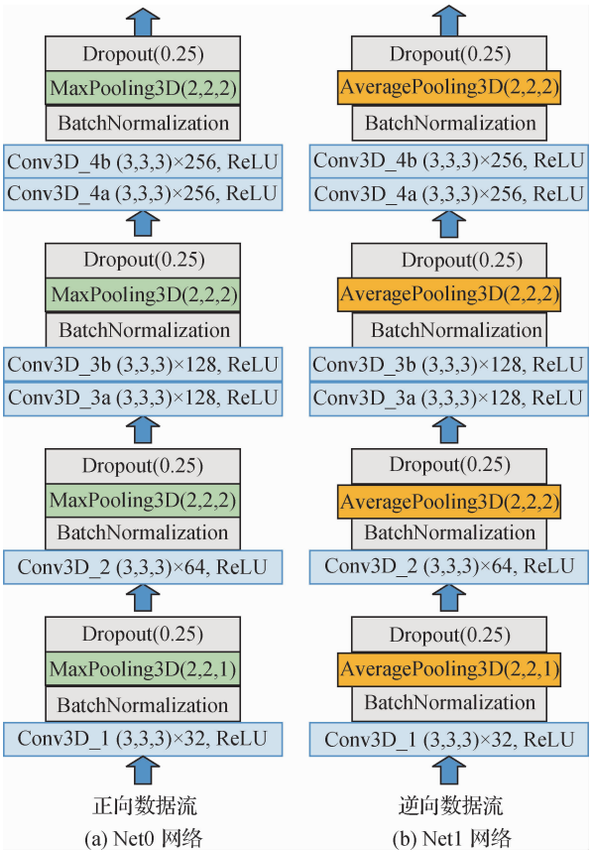


图 4　双路特征提取模块

Fig. 4　Two-way feature extraction module

((a) Net0 network; (b) Net1 network)

层的内核大小为 2×2×2,步长为2×2×2,使得在空间大小和时间长度上缩小比例为 4 和 2,同时在每级池化层后加入随机失活层(dropout)缓解模型过拟合。使用 BN 层通过规范化手段,让每层神经网络的输入值符合标准正态分布,这样避免了反向传播过程中的梯度消失,极大提升了训练效率,提高了分类精度,同时降低了模型中超参数的调节难度。

Net0 和 Net1 网络的不同之处在于视频输入层和池化层。Net0 网络接收数据时,按原始数据、两帧融合数据和三帧融合数据的顺序输入,采用最大池化(maxpooling3D),目的主要是保留纹理特征。而 Net1 网络接收数据时,按 Net0 网络的逆序输入,采用平均池化层(averagepooling3D),目的是保留整体的数据特征。

4　BiConvLSTM 网络

经典的 LSTM 网络由 input gate、forget gate、cell、output gate 和 hidden 等 5 个模块组成,这种结构适用于处理时间序列数据,但是处理图像这种空间结构数据,将会带来空间冗余,原因是空间数据具有很强的局部相关性,这种结构会忽略空间局部特征。为此,Shi 等人(2015)提出了 ConvLSTM 模型,有效解决了状态转换过程中空间特征丢失的问题。ConvLSTM 模型将输入与各个门之间的连接替换成了卷积,同时状态与状态之间也换成卷积运算,这样做的好处是在提取空间特征的同时能够建立时序关系。ConvLSTM 模型的工作原理为

$$\begin{aligned}
\boldsymbol{i}_t &= \delta(\boldsymbol{W}_{xi} * \boldsymbol{X}_t + \boldsymbol{W}_{hi} * \boldsymbol{H}_{t-1} + \boldsymbol{W}_{ci} \circ \boldsymbol{C}_{t-1} + b_i)\\
\boldsymbol{f}_t &= \delta(\boldsymbol{W}_{xf} * x_t + \boldsymbol{W}_{hf} * \boldsymbol{H}_{t-1} + \boldsymbol{W}_{cf} \circ \boldsymbol{C}_{t-1} + b_f)\\
\boldsymbol{C}_t &= \boldsymbol{f}_t \circ \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \circ \tanh(\boldsymbol{W}_{xc} * \boldsymbol{X}_t + \boldsymbol{W}_{hc} * \boldsymbol{H}_{t-1} + b_c)\\
\boldsymbol{o}_t &= \delta(\boldsymbol{W}_{xo} * \boldsymbol{X}_t + \boldsymbol{W}_{ho} * \boldsymbol{H}_{t-1} + \boldsymbol{W}_{co} \circ \boldsymbol{C}_t + b_0)\\
\boldsymbol{H}_t &= \boldsymbol{o}_t \circ \tanh(\boldsymbol{C}_t) \qquad (7)
\end{aligned}$$

式中,$\boldsymbol{X}_t$ 表示当前时刻的输入,$\boldsymbol{H}_{t-1}$ 表示 $t-1$ 时刻的输出,$*$ 表示卷积,$\circ$ 表示哈达玛积。i_t、f_t 和 o_t 分别表示输入门、遗忘门和输出门,$\boldsymbol{W}$ 表示遗忘门的递归权重,b 为偏置,$\boldsymbol{X}$、$\boldsymbol{C}$、$\boldsymbol{H}$、$\boldsymbol{i}$、$\boldsymbol{f}$、$\boldsymbol{o}$ 为 3 维张量。

参照此结构,设计了双向卷积长短期记忆网络(BiConvLSTM),结构如图 5 所示,采用两级 ConvLSTM 结构,第一层的输入为双路特征提取模块提取的特征的融合,两级 ConvLSTM 层的卷积滤波器分别为 128 和 256,卷积核的大小都为 3×3,输出序列的最后一个输出到 Softmax 分类器。为了增加网络的容量,可以增加网络层数,但是级数过深会导致过拟合,因此根据本文数据集的大小采用了两级结构。在 ConvLSTM 网络中嵌入双向层,将相同的信息以正序和逆序的方式呈现给网络,能捕捉到仅使用正序时可能忽略的一些模式,从而得到更加丰富的表示,提高精度并缓解遗忘问题,因此双向 ConvLSTM 网络能够更好地捕捉到视频的全局信息,获得更好的预测效果。

5　测试与分析

为了验证本文算法的有效性,采用行为识别研究中常用的 UCF101 和 HMDB51 数据集进行验证。UCF101 数据集分为 5 类:人与物体互动、人体动作、人与人互动、乐器演奏和体育运动,共 13 320 个视频。HMDB51 数据集包含面部动作、肢体动作、人类

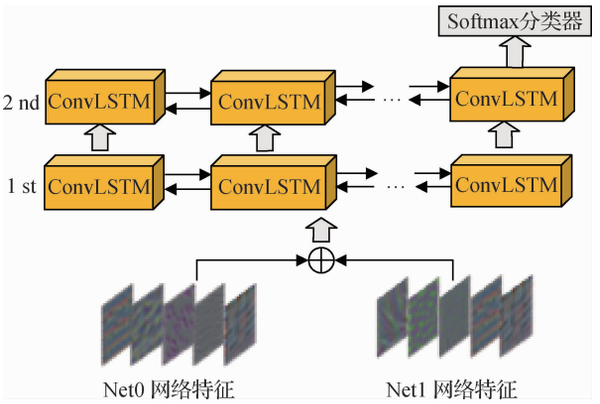


图 5　BiConvLSTM 模块

Fig. 5　BiConvLSTM module

交互之间的肢体动作等，共 6 766 个视频。模型的训练与测试采用英特尔 Core i7-8700k CPU，NVIDIA GTX2080Ti 显卡，12 GB 内存，以 python 语言为基础，在 Keras 和 Tensorflow 模型框架下编程实现。

5.1　数据预处理与参数设置

将原始视频剪裁为 80 × 80 像素分辨率的视频片段，通过时域融合模块中的两帧融合算法和三帧融合算法将长视频压缩成短视频，采用任何算法融合视频时，都是从原始数据中随机抽取互不重复的视频片段。最后，将原始数据集、两帧融合数据集以及三帧融合数据集以长度为 8 帧的视频片段作为网络的输入。

实验过程中采用小批量数据进行训练，每 16 批处理一次，迭代次数 epoch = 100。通过自动调整学习率的方式提高模型的准确率，调整过程采用线性衰减的方式，具体为

$$lr = lr \times \alpha \tag{8}$$

学习率的下限为 $lr_{\min} = 0.000\ 1$，初始学习率 $lr = 0.01$，α 为学习率下降因子，当学习停滞时，如果在 3 个 epoch 中检测不到模型性能提升，则按式(9)减少学习率，其中 $\alpha = 0.5$。实验过程中的参数设置如表 1 所示。

表 1　参数设置

Table 1　Parameter settings

参数名	UCF101	HMDB51
数据集形状	(39 960, 80, 80, 8, 3)	(20 328, 80, 80, 8, 3)
训练集形状	(31 968, 80, 80, 8, 3)	(16 262, 80, 80, 8, 3)
验证集形状	(7 992, 80, 80, 8, 3)	(4 066, 80, 80, 8, 3)
初始学习率	$lr = 0.01$	
学习率下降因子	$\alpha = 0.5$	
优化函数	Optimizer = “SGD”	
损失函数	loss = categorical_crossentropy，交叉熵损失函数	
迭代次数	epoch = 100	
小批量数据	bacth_size = 16	
权重衰减	weight_decay = 0.006	

5.2　交叉验证

为了更好地评估本文模型的抗干扰能力和泛化能力，通过计算 5 折交叉验证的平均准确率来评判模型，过程如下：

1)将原始数据集、两帧融合数据集和三帧融合数据集打乱，分为 5 个部分，使其互不相交。

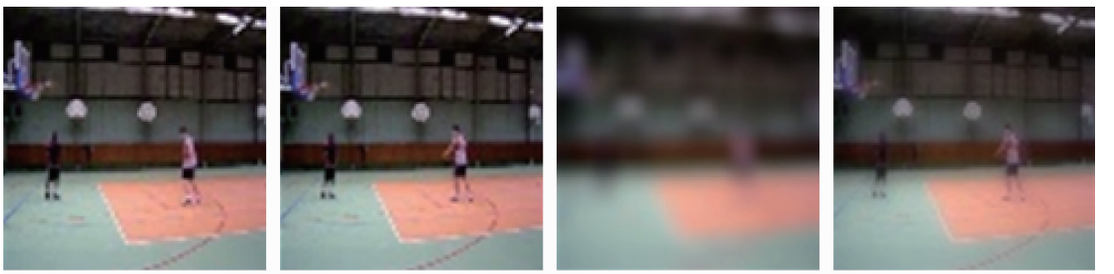
2)每次取出其中的 1 个部分作为测试集，将其余部分数据的 80% 作为训练集，20% 作为验证集。

3)将训练 5 次的准确率求平均值，最终得到模型的准确率。

5.3　时域融合模块测试

图 6 为两帧图像的融合过程。图 6(a)和图 6(b)分别为 UCF101 数据集中随机抽取的两帧图像，图 6(c)为图 6(a)的低频信息，图 6(d)为图 6(b)和图 6(c)的融合图。

图 7 为三帧图像的融合过程，图 7(a)—(c)分别为从 UCF101 数据集中随机抽取的 3 帧图像，图 7(d)为图7(a)的低频信息，图7(e)为图7(b)的低



(a) $f_{k-1}(x,y)$ 帧　(b) $f_k(x,y)$ 帧　(c) $L(x,y)$ 帧　(d) $F(x,y)$ 帧

图 6　两帧融合过程

Fig. 6　Two frame fusion ((a) $f_{k-1}(x,y)$; (b) $f_k(x,y)$; (c) $L(x,y)$; (d) $F(x,y)$)

(a) $f_{k-1}(x,y)$ 帧　(b) $f_k(x,y)$ 帧　(c) $f_{k+1}(x,y)$ 帧　(d) $L_{k-1}(x,y)$ 帧　(e) $L_k(x,y)$ 帧　(f) $M(x,y)$ 帧

图 7　三帧融合过程

Fig. 7　Three frame fusion ((a) $f_{k-1}(x,y)$; (b) $f_k(x,y)$; (c) $f_{k+1}(x,y)$; (d) $L_{k-1}(x,y)$; (e) $L_k(x,y)$; (f) $M(x,y)$)

频信息,图 7(f)为图 7(d)、图 7(e)和图 7(b)的融合图。

低频信息提取参数设置如表 2 所示。

表 2　融合模块参数设置

Table 2　Fusion module parameter settings

融合方式	方差	高斯核
两帧融合	δ = (3,3)	kernel_size = (9,9)
三帧融合	δ = (3,3)	kernel_size = (5,5)

数据集中的所有短视频按两帧和三帧算法融合,融合后的短视频中不但含有源视频图像的大部分信息,而且降低了原始视频的数据量。从理论上讲,输入网络 8 帧的视频片段,在两帧融合算法下包含了原视频 16 帧图像的大部分信息,在三帧融合算法下包含了 24 帧原始视频,这种处理降低了网络设计的复杂度。

5.4　测试与分析

为了验证本文提出的融合方法能否提高网络对视频信息的分析和提取能力,以及本文模型的泛化能力和稳定性,采用两种方式对模型进行评估。1)采用 5 折交叉验证的方式对行为进行整体测试与分析;2)针对每类行为动作进行比较统计,分析每类行为的准确率。在公开的行为识别数据集 UCF101 和 HMDB51 上进行 5 折交叉验证,在验证集上的平均准确率分别为 96.47% 和 80.03%,如表 3 所示。

表 3　交叉验证的准确率

Table 3　Cross validation accuracy

/%

数据集	实验次数					平均
	Split_1	Split_2	Split_3	Split_4	Split_5	
UCF101	96.52	96.40	96.22	96.63	96.57	96.47
HMDB51	81.03	80.39	78.29	80.11	80.31	80.03

图 8 展示了 UCF101 和 HMDB51 数据集中每类行为的准确率。可以看出,在 UCF101 数据集上,几乎所有类别的分类精度都大致相等,接近平均值,仅行为类别 cliffdiving 上的准确率明显低于平均值,主要原因是该行为动作背景的复杂性和多样性造成的。在 HMDB51 数据集上,不同类别的分类精度相差较大,主要原因是数据集中含有大量交互方面的视频动作,视频复杂度大。

3D ConvNets 封装了视频中与目标、场景、动作有关的信息,这些特征在行为分类任务中具有重要作用,是一种高效的视频描述符,能够从中较好地提

取短时空特征。为了提取视频中的细节特征和整体特征，将融合数据顺序输入 Net0 网络提取其细节特征，将融合数据逆向输入 Net1 网络提取整体特征，最后将两路特征融合，完成行为动作的特征建模。BiConvLSTM 网络处理视频数据时，对每一帧图像提取空间特征，然后依次将视频帧建立时序关系，不仅可以像 LSTM 一样建立时序关系，还可以像卷积神经网络一样刻画局部空间特征。为了对时间序列中间隔和延迟相对较长的行为类别建模，设计了 BiConvLSTM 网络对双路融合特征层建模，进一步提取了局部特征并建立时序关系，很好地融合了两种网络的优点，因此本文模型具有很高的准确率。

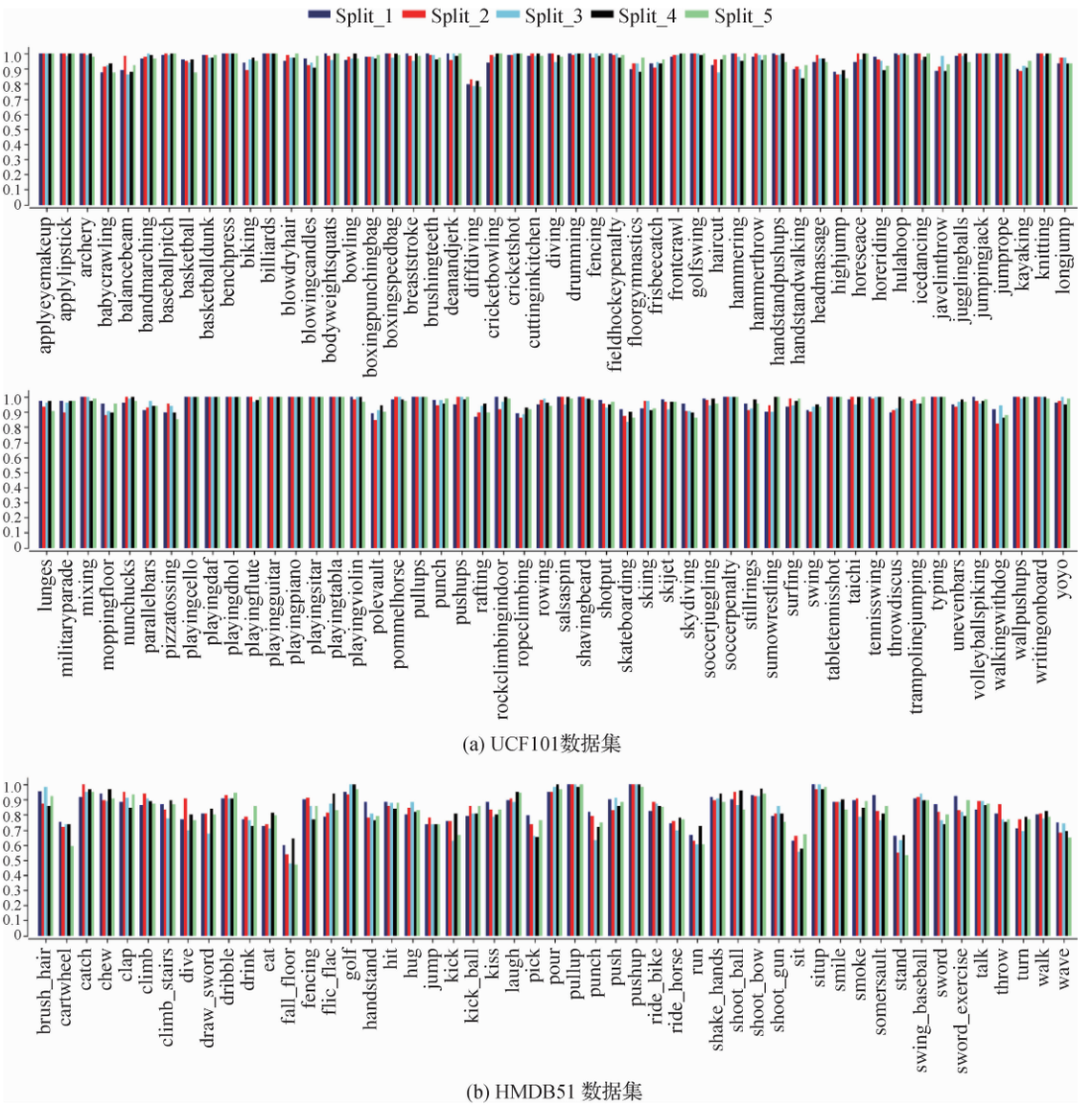


图 8　UCF101 和 HMDB51 数据集上每类行为的准确率分析

Fig. 8　Analysis of accuracy of each type behavior on UCF101 and HMDB51 datasets

((a) UCF101 datasets; (b) HMDB51 datasets)

5.5　本文方法与目前典型主流模型对比

为了评价本文算法的性能，与传统算法、双流结构、3D 卷积神经网络、LSTM 结构、融合结构进行比较，结果如表 4 所示。实验测试了各算法在 UCF101 和 HMDB51 数据集上 5 折交叉下的平均识别精度。在 UCF101 数据集上，本文方法优于其他算法，与经典的双流网络(Simonyan 和 Zisserman，2014)和最新的融合算法(Ouyang 等，2019)相比，准确率分别提高了 12.55% 和 3.15% 。在 HMDB51 数据集上，本文算法也获得了最好的识别精度，与最新的算法

Multi-task C3D + LSTM(Crasto 等,2019)相比,准确率提高了 2.12%。

表 4 本文模型与目前典型模型在 UCF101 和 HMDB51 数据集上的准确率对比

Table 4 Comparison of the accuracy of the model in this article and the current typical model on UCF101 and HMDB51 datasets

/%

模型	UCF101	HMDB51
iDT(Wang 和 Schmid,2013)	86.40	57.20
Two-Stream(Simonyan 和 Zisserman,2014)	88.00	59.40
C3D(Tran 等,2015)	82.30	56.80
RGB + RGBF + IDT(Yang 等,2019)	92.60	65.40
Composite LSTM(Li 等,2017)	89.20	56.40
(C3D + iDT) - SVM(Tran 等,2015)	90.40	-
TSN(Wang 等,2016)	94.20	69.50
Spatiotemporal Heterogeneou(Chen 等,2019)	94.40	67.20
Multi-task C3D + LSTM(Ouyang 等,2019)	93.40	68.90
Motion-Augmented RGB Stream(Crasto 等,2019)	-	79.50
S-TPNet(Zheng 等,2019)	-	74.80
本文	**96.47**	**80.03**

注:加粗字体表示各列最优结果,"-"表示未在数据集上测试。

本文方法具有优良的识别精度,缘于以下几个方面:1)基于手动提取特征的方法,准确率由手动特征限制。手动特征提取好,则识别率高,反之,识别率低。对于海量数据来说,手动特征提取速度慢,特征很难完全描述行为。2)基于双流结构的方法通过双路 CNN 分别提取图像帧和光流帧的特征并融合分类。识别率相对传统方法有很大的提高,但是丢失了动作的时序关联信息,同时光流特征对背景不变的视频表述比较好。3)3D 卷积神经网络可以同时对外观和运动信息进行建模,在行为分类中都优于 2D 卷积网络,但是受到数据集的限制以及参数量等因素的影响,行为准确率较低。4)LSTM 结构主要对提取的 3D 特征或者 2D 特征按时间轴建模,提取了视频数据的时序特征,但是丢失了局部空间特征。

本文方法在 UCF101 和 HMDB51 数据集上提升很明显,主要是通过 3 个过程提高了模型的准确率。1)在时域上通过两帧和三帧融合提取原始视频的信息,提高对视频信息的分析能力,让其能够完整表达行为信息;2)在网络结构上通过双路特征提取网络,提取行为的细节特征和整体特征,高效提取了时空特征;3)对融合特征进一步提取细节特征并建立时序关系,解决了视频时间序列中某些行为的执行方式不同、间隔相对较长等因素对行为识别的影响。

6 结 论

本文采用 3D ConvNets 和 BiConvLSTM 基础网络,提出了多特征融合的行为识别模型。首先对行为数据集中的低频信息和高频信息加权融合,压缩数据集中的冗余信息,增强行为识别数据集,然后设计了双路特征提取网络,分别提取行为动作的细节特征和整体特征并建模,最后验证了本文模型的性能。实验对比了本文模型与传统算法、双流结构、3D 卷积神经网络、LSTM 结构和融合结构在行为识别数据集 UCF101 和 HMDB51 上的准确率,本文模型在验证集上分别达到了 96.47% 和 80.03% 的分类精度,有效提高了行为识别的准确率。本文模型可以在智能安防中辅助人工审核监控视频中出现的特定行为。

未来将从两个方面进一步深入研究多特征融合的行为识别模型。1)从特征提取的角度出发,参照

ResNet、GoogLeNet 和 DenseNet 等经典的卷积神经网络，设计 3D 特征提取网络，高效提取空间和时间特征。2）从时域融合的角度出发，设计视频融合算法，去除原始数据中的冗余信息，高效表示行为特征。

参考文献（References）

Chen E Q, Bai X, Gao L, Tinega H C and Ding Y Q. 2019. A spatio-temporal heterogeneous two-stream network for action recognition. IEEE Access, 7: 57267-57275 [DOI: 10.1109/access.2019.2910604]

Crasto N, Weinzaepfel P, Alahari K and Schmid C. 2019. MARS: motion-augmented RGB stream for action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7874-7883 [DOI: 10.1109/cvpr.2019.00807]

Diba A, Fayyaz M, Sharma V, Karami A H, Arzani M M, Yousefzadeh R and van Gool L. 2017. Temporal 3d convnets: new architecture and transfer learning for video classification[EB/OL]. [2017-11-22]. https://arxiv.org/pdf/1711.08200.pdf

Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T and Saenko K. 2015. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE: 2625-2634 [DOI: 10.1109/CVPR.2015.7298878]

Harris C and Stephens M. 1988. A combined corner and edge detector//Proceedings of Alvey Vision Conference. Manchester, UK: Alvey Vision Club: 147-151 [DOI: 10.5244/c.2.23]

Ji J W, Buch S, Soto A and Niebles J C. 2018. End-to-end joint semantic segmentation of actors and actions in video//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 734-749 [DOI: 10.1007/978-3-030-01225-0_43]

Kläser A, Marszalek M and Schmid C. 2008. A spatio-temporal descriptor based on 3d-gradients//Proceedings of British Machine Vision Conference 2008. Leeds, UK: British Machine Vision Association: 275-284 [DOI: 10.5244/c.22.99]

Laptev I and Lindeberg T. 2003. Space-time interest points//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE: 432-439 [DOI: 10.1109/iccv.2003.1238378]

Li Y G, Ge R, Ji Y, Gong S R and Liu C P. 2019. Trajectory-pooled spatial-temporal architecture of deep convolutional neural networks for video event detection. IEEE Transactions on Circuits and Systems for Video Technology, 29(9): 2683-2692 [DOI: 10.1109/tcsvt.2017.2759299]

Li Z Y, Gavrilyuk K, Gavves E, Jain M and Snoek C G M. 2018. Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding, 166: 41-50 [DOI: 10.1016/j.cviu.2017.10.011]

Liu J, Shahroudy A, Xu D and Wang G. 2016. Spatio-temporal LSTM with trust gates for 3D human action recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 816-833 [DOI: 10.1007/978-3-319-46487-9_50]

Lowe D G. 1999. Object recognition from local scale-invariant features//Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE: 1150-1157 [DOI: 10.1109/iccv.1999.790410]

Ma Y X, Tan L, Dong X and Yu C C. 2019. Action recognition for intelligent monitoring. Journal of Image and Graphics, 24(2): 282-290（马钰锡，谭励，董旭，于重重. 2019. 面向智能监控的行为识别. 中国图象图形学报，24(2)：282-290）[DOI: 10.11834/jig.180392]

Ng J Y H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R and Toderici G. 2015. Beyond short snippets: deep networks for video classification//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4694-4702 [DOI: 10.1109/cvpr.2015.7299101]

Ouyang X, Xu S J, Zhang C Y, Zhou P, Yang Y, Liu G H and Li X L. 2019. A 3D-CNN and LSTM based multi-task learning architecture for action recognition. IEEE Access, 7: 40757-40770 [DOI: 10.1109/access.2019.2906654]

Qiu Z F, Yao T and Mei T. 2017. Learning spatio-temporal representation with pseudo-3D residual networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5534-5542 [DOI: 10.1109/ICCV.2017.590]

Scovanner P, Ali S and Shah M. 2007. A 3-dimensional sift descriptor and its application to action recognition//Proceedings of the 15th ACM International Conference on Multimedia. Augsburg, Germany: ACM: 357-360 [DOI: 10.1145/1291233.1291311]

Sharma S, Kiros R and Salakhutdinov R. 2015. Action recognition using visual attention[EB/OL]. [2016-02-14]. https://arxiv.org/pdf/1511.04119.pdf

Shi X J, Chen Z R, Wang H, Yeung D Y, Wong W K and Woo W C. 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[DB/OL]. [2018-09-30]. https://www.researchgate.net/publication/278413880_Convolutional_LSTM_Network_A_Machine_Learning_Approach_for_Precipitation_Nowcasting

Simonyan K and Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS: 568-576

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Pro-

ceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/iccv.2015.510]

Wang H, Kläser A, Schmid C and Liu C L. 2011. Action recognition by dense trajectories//Proceedings of CVPR 2011. Providence, USA: IEEE: 3169-3176 [DOI: 10.1109/cvpr.2011.5995407]

Wang H and Schmid C. 2013. Action recognition with improved trajectories//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 3551-3558 [DOI: 10.1109/iccv.2013.441]

Wang L, Koniusz P and Huynh D. 2019. Hallucinating IDT descriptors and I3D optical flow features for action recognition with CNNs//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8697-8707 [DOI: 10.1109/ICCV.2019.00879]

Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O and van Gool L. 2016. Temporal segment networks: towards good practices for deep action recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 20-36 [DOI: 10.1007/978-3-319-46484-8_2]

Wang X H, Gao L L, Wang P, Sun X S and Liu X L. 2018. Two-stream 3-D ConvNet fusion for action recognition in videos with arbitrary size and length. IEEE Transactions on Multimedia, 20(3): 634-644 [DOI: 10.1109/tmm.2017.2749159]

Willems G, Tuytelaars T and van Gool L. 2008. An efficient dense and scale-invariant spatio-temporal interest point detector//Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer: 650-663 [DOI: 10.1007/978-3-540-88688-4_48]

Yang H, Yuan C F, Li B, Du Y, Xing J L, Hu W M and Maybank S J. 2019. Asymmetric 3D convolutional neural networks for action recognition. Pattern Recognition, 85: 1-12 [DOI: 10.1016/j.patcog.2018.07.028]

Zheng Z X, An G Y, Wu D P and Ruan Q Q. 2019. Spatial-temporal pyramid based Convolutional Neural Network for action recognition. Neurocomputing, 358: 446-455 [DOI: 10.1016/j.neucom.2019.05.058]

Zhu W J, Hu J, Sun G, Cao X D and Qiao Y. 2016. A key volume mining deep framework for action recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1991-1999 [DOI: 10.1109/cvpr.2016.219]

作者简介

谭等泰,1986 年生,男,讲师, 主要研究方向为机器学习和行为识别。
E-mail: 465402383@qq.com

李世超,男,副教授,主要研究方向为 5G 无线通信和无线资源管理。E-mail: 739998452@qq.com

常文文,男,博士,主要研究方向为信号处理、模式识别和机器学习。E-mail: changww2013@126.com

李登楼,女,讲师, 主要研究方向为物证技术和行为识别。
E-mail: 964779559@qq.com