

引用格式:王震,刘瑞敏,黄琼桃. 用于人体行为识别的 Inflated VGGNet-16 网络[J]. 北京化工大学学报(自然科学版),2020,47(3):114-121.

WANG Zhen, LIU RuiMin, HUANG QiongTao. Inflated VGGNet-16 networks for human action recognition[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2020,47(3):114-121.

# 用于人体行为识别的 Inflated VGGNet-16 网络

王 震 刘瑞敏\* 黄琼桃

(昆明理工大学 信息工程与自动化学院, 昆明 650504)

**摘 要:** 针对目前人体行为识别算法中 C3D 网络结构较浅、特征提取能力差,以及无可用预训练模型、训练耗时长等问题,以更深的 VGGNet-16 网络为基础,通过添加批归一化层(batch normalization layer)以及使用 Inflating 方法将 ImageNet 预训练模型用于网络初始化,设计了一种新型的人体行为识别 3D 网络。通过在标准数据集 UCF101 与 HMDB-51 上的实验分析,将图片进行中心剪切后作为所设计网络的输入,从零训练时在 UCF101 数据集上比原始 C3D 网络的精度提高了 9.2%,并且网络收敛速度更快,验证了所设计的 Inflated VGGNet-16 网络具有更强的特征提取与泛化能力。最后,将所设计网络加上 10 倍数据增强,在两个标准数据集上准确率分别达到了 89.6% 与 61.7%,相比于较浅的 C3D 网络在 UCF101 数据集上提升了 7.3%,超过了传统的改进密集轨迹法(iDT)以及经典的双流卷积神经网络(two-stream),具有较高的行为识别准确率。

**关键词:** 行为识别; VGGNet-16; Inflating; ImageNet 预训练; 数据增强

**中图分类号:** TP399 **DOI:** 10.13543/j.bhxbzr.2020.03.015

## 引 言

人体行为识别也称为行为分类,是指利用模式识别、机器学习等方法,将一段视频中的人体行为分类到预先定义的几种行为类别中。人体行为特征提取分为手工提取与自动学习<sup>[1-2]</sup>。在深度学习占据该领域主导地位之前,基于手动特征的改进密集轨迹法(improved dense trajectories, iDT)是最典型也是识别精度最高的算法<sup>[3-4]</sup>。该方法通过对图片进行密集采样并在一小段视频序列中对采样点进行跟踪,进而得到用于提取特征的密集轨迹,最后将特征编码利用支持向量机(support vector machine, SVM)进行分类。iDT 算法有着不错的分类效果和很好的鲁棒性,但效率不高、速度慢。之后 Simonyan 等<sup>[5]</sup>提出了分别以单帧 RGB 图片与堆叠光流图为网络输入的双流卷积神经网络(two-stream),发现基于深度学习的方法在识别精度上超过了传统的基于手动

特征的 iDT 算法,双流卷积神经网络通过模拟人体视觉神经机制达到当时最高的识别准确率。随后 Feichtenhofer 等<sup>[6]</sup>提出在双流中进行多组融合实验可以进一步提高识别准确率。Donahue 等<sup>[7]</sup>针对双流卷积神经网络缺乏对视频长时信息建模问题,提出 convolutional neural networks + long short-term memory (CNN + LSTM) 的结构,从视频中抽取若干帧经 CNN 提取特征后,输入到递归神经网络提取时域信息,最后进行分类。虽然该结构能够对时间序列建模,但每一帧提取特征之后丢失了底层的时间信息。Tran 等<sup>[8]</sup>提出了直接利用连续的视频帧作为输入的 C3D 网络结构,3D 网络比 2D 网络具有更快的处理速度以及更好的特征表示,但其性能受限于数据集的规模。

近几年的研究多为 C3D 与双流卷积神经网络,并结合 iDT 方法提高网络识别精度<sup>[9-10]</sup>。对于多数双流卷积神经网络<sup>[11-12]</sup>,其时间流以堆叠的光流图作为网络输入,计算量巨大,耗时几乎占整个网络训练的 90%,因此实时性差严重制约了其在现实生活中的应用。尽管也有人提出 dynamic image<sup>[13-14]</sup>、motion vector<sup>[15]</sup>、RGB Diff<sup>[16]</sup> 等替代光流作为视频运动表征的方法,但效果都没有光流好。

收稿日期: 2019-12-03

基金项目: 国家自然科学基金(61863018)

第一作者: 男,1995 年生,硕士生

\* 通信联系人

E-mail: 2500013476@qq.com

对于3D卷积网络,原始的C3D深度比较浅,特征提取能力较弱,未能真正发挥其性能,其后研究中的3D网络大都基于复杂的大型网络结构,如Inflated 3D(I3D)<sup>[17]</sup>、Pseudo-3D(P3D)<sup>[18]</sup>、 $R(2+1)D$ <sup>[19]</sup>、T-C3D<sup>[20]</sup>等。其中P3D、 $R(2+1)D$ 与T-C3D都是将二维卷积神经网络中的改进方法成功运用到了三维卷积神经网络,对网络的识别准确率提升较小,而I3D将网络从二维扩展到三维,利用二维网络预训练模型对三维模型初始化,提高网络的收敛速度,降低过拟合及网络训练对数据量的要求,在UCF101与HMDB-51数据集上达到了目前最高的准确率,分别为98.0%、80.9%<sup>[21-22]</sup>。除了对网络结构优化外,Hara等<sup>[23]</sup>通过实验证明了大型视频数据集Kinetics可以满足151甚至200层ResNet 3D CNN的训练;Köpkülü等<sup>[24]</sup>则对不同复杂程度的3D网络分析,指出不应为了节省复杂度而将3D CNN设计的太浅或太窄;Tran等<sup>[25]</sup>将分组卷积引入视频分类任务并验证了通道交互数量在3D组卷积网络中对精确度有着重要的作用;Crasto等<sup>[26]</sup>提出motion-augmented RGB stream(MARS)以避免在测试时计算运动信息,提高网络的实用性,但在训练过程中仍然需要计算运动信息。以上这些网络的训练需要消耗很大的计算机资源,且在基础的C3D网络上并没有进行改进与进一步研究。因此本文针对以上问题,以VGGNet-16替换C3D中较浅的VGGNet-A网络为基础,利用Inflating方法将其从2D网络扩展为3D网络并添加批标准化层(batch normalization),最后使用ImageNet上的预训练模型对其进行初始化,

并结合10倍数据增强设计了一种用于人体行为识别的Inflated VGGNet-16网络。

## 1 人体行为识别及常用数据集

人体行为识别在安全监控、室内医护、机器人、自动驾驶、人机交互等众多领域越来越受到关注,也成为了计算机视觉研究者们越来越流行的主题,其任务是从一段未知的视频中自动分析出人体发生的行为。从视频中发生动作的人物数量上可将其分为单人识别与多人识别,从输入视频是否经过修剪可以分为修剪视频的行为识别和未修剪视频的行为识别,后者是ActivityNet大赛<sup>[27]</sup>上的两项重要的行为识别挑战项目。在人体行为识别领域最常用的标准数据集是UCF101与HMDB-51数据集。UCF101来源于YouTube,主要包括人物交互、肢体动作、人人交互、奏乐器和各类运动共5大类动作的101种类别,其中每类动作由25个人完成,每人做4~7组,共13 320段视频,分辨率为 $342 \times 256$ ,是目前包含动作种类最多的数据集。该数据集在动作的采集上具有非常大的多样性,包括相机运行、外观变化、姿态变化、物体比例变化、背景变化、光线变化等。HMDB-51数据集由Brown University发布,视频多数来源于电影,还有一部分来自公共数据库以及YouTube等网络视频库,包含有6 849段样本,主要包含面部动作、肢体动作和人与人之间肢体动作三大类别共51种动作类,每类动作至少包含有101段样本,分辨率为 $320 \times 240$ 。两个数据集的部分动作示例如图1所示,图1(a)为UCF101数据集中的拖地板、剃胡子、打拳击、冲浪与



图1 UCF101与HMDB-51部分动作示例

Fig. 1 UCF101 and HMDB-51 parts of the example action

跳绳动作,图 1(b)为 HMDB-51 数据集中的攀岩、跑、体操、足球与射击动作。

2 Inflating 方法与 C3D 网络

Inflating 方法来源于 Carreira 等<sup>[17]</sup>提出的 I3D 网络,该方法可以将 2D 的预训练模型在时间维上扩张用于 3D 网络初始化,不仅有效利用了 Image Net 数据集上的预训练模型,而且可以使网络从大型数据集繁重的训练过程中脱离出来,很好地应用于各种小型数据集,加速网络训练、提高分类准确度,其示意图如图 2 所示。其扩张方法是将 2D 卷积核在时间维度复制  $N$  次,然后再将所有的权重除以  $N$  得到 3D 卷积核,模型的偏置与 2D 网络相同,对于网络中的其他层则直接转化为 3D。

原始的 C3D 网络结构如图 3(a)所示,包含 8 个卷积层,5 个最大池化层和 2 个全连接层,最后是

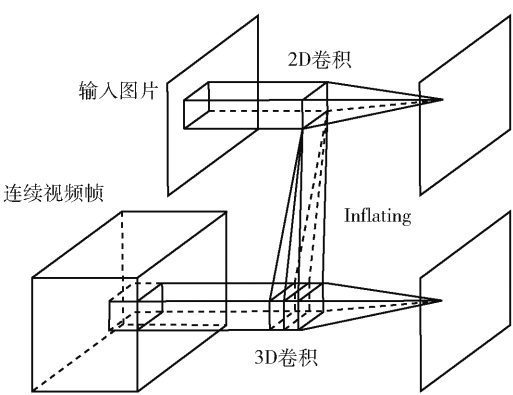
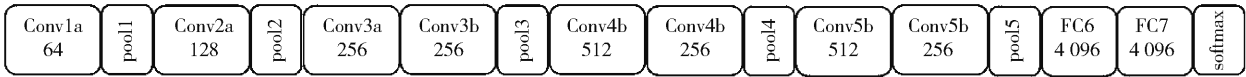
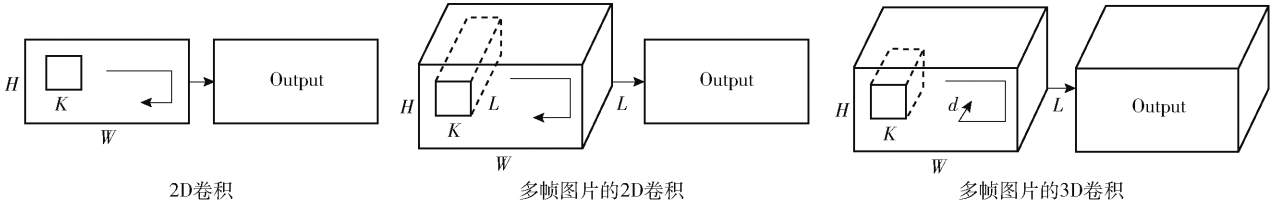


图 2 Inflating 方法示意图

Fig. 2 Schematic diagram of the Inflating method softmax 输出层。图 3(b)为 2D 卷积与 3D 卷积示意图,可以看出不管是输入单张图片还是多帧图片,通过 2D 卷积后输出都为单通道响应图,而多帧图片经过 3D 卷积后,输出为多通道响应图,可以很好地保留时间信息。



(a) C3D网络结构



(b) 卷积操作

图 3 C3D 网络结构与卷积操作

Fig. 3 C3D network structure and convolution operations

3 网络设计与数据处理

3.1 网络结构

本文网络以二维 VGGNet-16 结构<sup>[28]</sup>为基础,通过增加时域维度将其扩展为三维网络结构,该网络中卷积核尺寸与原 C3D 网络一致,都为  $3 \times 3 \times 3$ ,时域与空间步长均为 1,池化层都采用最大池化,为了不在浅层破坏输入数据的时域信息,第一层池化层尺寸为  $1 \times 2 \times 2$ ,其余为  $2 \times 2 \times 2$ ,步长为 2,激活函数选用线性修正单元 (rectified linear unit, ReLU)<sup>[8]</sup>。整个网络结构除了在深度上有所增加外,另一个重要的改进为在所有的卷积层以及全连接层后加入了批标准化层<sup>[29]</sup>。网络结构如图 4 所示,图中共有 5 组卷积,数字表示卷积核的数目,最后全连接层均接 Dropout 层,随机失活率为 0.5。

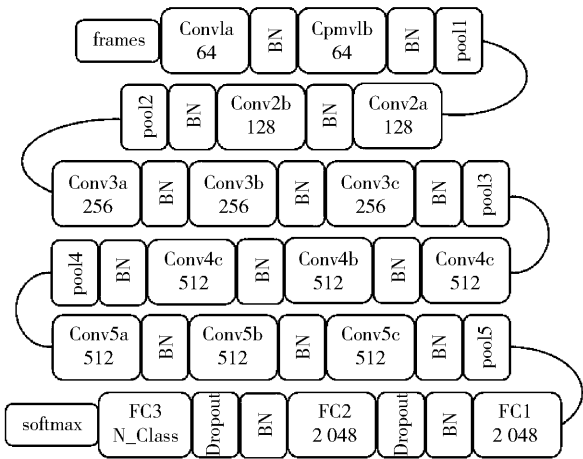


图 4 Inflated VGGNet-16 网络结构

Fig. 4 Inflated VGGNet-16 network structure

3.2 批标准化

神经网络的训练与测试集应具有相同的分布。



对于深度学习这种包含很多隐藏层的网络结构,在训练过程中各层的参数不停地改变,每个隐藏层都会面临协变量偏移的情况,也就是说,隐藏层的输入数据分布总是来回改变,这非常不利于网络对数据的学习。批标准化层的作用就是使每一隐藏层的输入保持相同分布,其前向传导流程如下。

给定一批数据  $B = \{x_{1,2,\dots,m}\}$ , 可学习参数  $\gamma$  和  $\beta$ , 首先利用式(1)计算每一个训练批次数据的均值。

$$\mu_B = \frac{1}{m} \sum_{i=1}^m x_i \quad (1)$$

然后利用式(2)计算该批次数据的方差。

$$\sigma_B^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2 \quad (2)$$

之后使用求得的均值和方差利用式(3)对该批次的训练数据作归一化,获得 0-1 分布,其中  $\varepsilon$  是为了避免除数为 0 而使用的微小正数。

$$x'_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \quad (3)$$

最后利用式(4)进行尺度变化和偏移,将  $x'_i$  乘以  $\gamma$  调整数值大小,再加上  $\beta$  增加偏移后得到  $y_i$ , 其中  $\gamma$  是尺度因子,  $\beta$  是平移因子,都为可学习参数。解决了因归一化后  $x_i$  被限制在正态分布下,出现网

络的表达能力下降的问题。

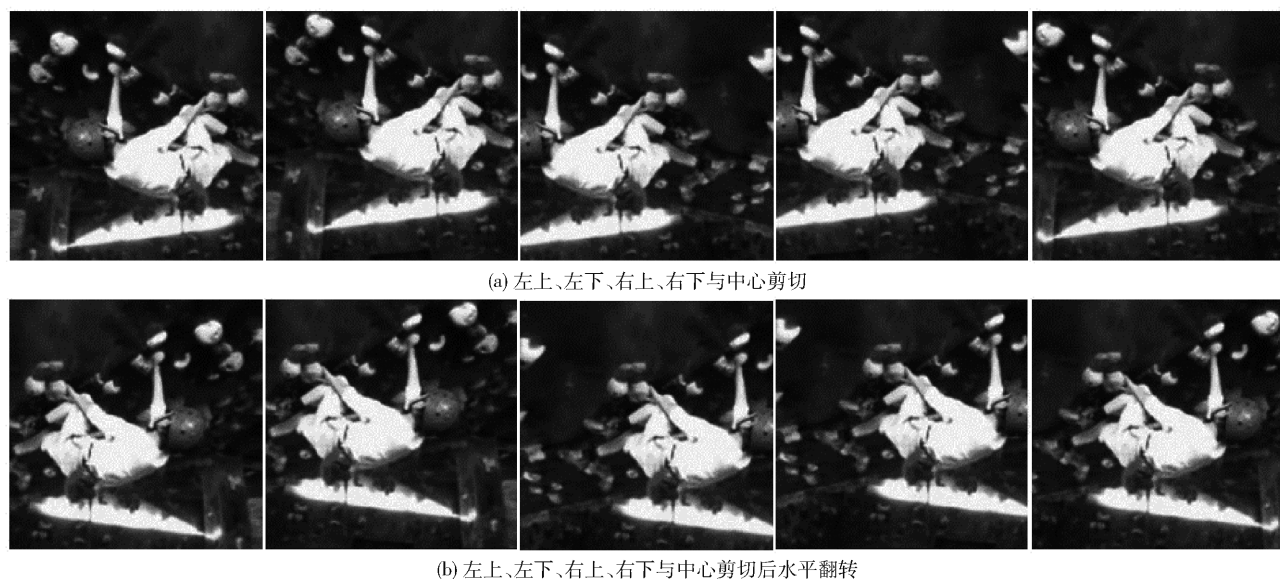
$$y_i = \gamma x'_i + \beta \quad (4)$$

批标准化不仅能够提高训练速度,还能防止过拟合,更重要的是降低了对超参数初始化的要求,可以使用大的学习率简化调参过程。

### 3.3 视频帧提取与数据增强

利用 open source computer vision library (OpenCV) 计算机视觉工具库对两个标准数据集进行视频帧的提取,UCF101 数据集中视频帧率为 25 帧/s,共包含约 248.6 万帧,占 29.3 G 内存。HM-DB-51 数据集帧率为 30 帧/s,共包含约 37.8 万帧,占 4.8 G 内存,所有帧都保存为 jpg 格式。

对于神经网络的训练,数据量不足一直都是影响其性能的重要因素,而 3D 网络的训练则需要比 2D 更多的数据,因此对其进行数据增强可以弥补这方面的不足。由于视频拍摄者一般都将运动目标置于视频的中心位置,因此为了使网络不仅仅关注于视频帧的中心部位,对其进行 corner cropping 即四角与中心剪切,然后随机水平翻转,使数据得到 10 倍增强,增强效果如图 5 所示,图 5(a) 从左到右为图 1 中攀岩动作的左上、左下、右上、右下四角与中心剪切,图 5(b) 为图 5(a) 中相应剪切后的水平翻转。



(a) 左上、左下、右上、右下与中心剪切

(b) 左上、左下、右上、右下与中心剪切后水平翻转

图5 数据增强

Fig.5 Data augmentation

## 4 实验结果与分析

在 Ubuntu 16.04 系统、RTX2070 显卡下以 Ten-

sorFlow 为后端的 Keras 框架进行实验。整个网络的超参数为:迭代最大轮次 16、批大小 12、输入图像维度(112,112,16,3)分别表示高、宽、连续帧数与通

道数,初始学习率 0.005,每经过 4 轮学习率降为原来的十分之一、优化器为随机梯度下降,使用 Nesterov 动量,动量值为 0.9。在测试时将输入视频进行时域分割<sup>[30]</sup>为 3 段,从每段视频中抽取 16 帧连续 RGB 图片作为网络输入,对于视频帧数小于要求的,对其进行循环复制,将 3 个视频段的得分平均值作为最终的分类结果。

表 1 为通过将数据进行中心剪切后作为输入时本文所设计 Inflated VGGNet-16 网络与 C3D 网络及其他主流网络的准确率对比,图 6 为表 1 中相应实验的数据曲线,其中 IVGG-16、19 与 IRes-18、34、50 分别表示 VGG-16、19 与 Res-18、34、50 对应的 Inflated 版本。

表 1 本文算法与 C3D 算法及其他主流网络在 UCF101 数据集 Split1 上的对比

Table 1 Comparison of the algorithm developed in this paper with the C3D algorithm and other Inflated base models in the UCF101 dataset Split1

方法	准确率/%
C3D	63.1
VGGNet-16(3D)	49.7
Inflated ResNet-18	43.2
Inflated ResNet-34	57.6
Inflated ResNet-50	61.8
Inflated VGGNet-16	72.3
Inflated VGGNet-19	70.9

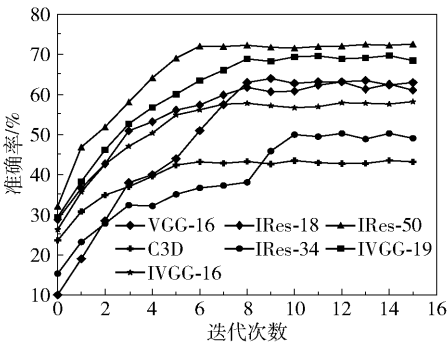


图 6 UCF101 Split1 验证集准确率曲线

Fig.6 Accuracy curves for UCF101 Split1 dataset validation

通过表 1 与图 6 可以看出小型数据集无法满足 VGGNet-19 深层网络模型的训练,而对于相对较浅的残差网络 ResNet-18 及 ResNet-34 又无法很好地对行为动作建模,在 VGGNet-16、VGGNet-19、ResNet-18、ResNet-34 及 ResNet-50 的各个扩展三维网

络中,Inflated VGGNet-16 网络识别性能最好、准确率最高。同时可以看出,简单地将 VGGNet-16 网络从 2D 扩展到 3D,从零开始训练时准确度不如 C3D 网络且收敛速度缓慢,其原因一方面是数据量不够,另一方面是网络从零初始化。然而将 ImageNet 数据集上预训练的模型扩展到 3D 并对其初始化之后,网络的收敛速度有所提高,并且精确度达到 72.3%。

表 2 为本文算法在两个数据集的 3 个标准训练-测试集分组方案(Split1、Split2、Split3)上的准确率,输入数据进行 10 倍增强,并在网络中加入批归一化层,算法的最终准确率取 3 个部分的平均值。

图 7、8 分别为在 UCF101 与 HMDB-51 上 3 个方案相应的测试准确率曲线。测试时批大小为 32,在两个数据集上随机选取 16 批进行测试,取平均值作为最终的准确率。

表 2 本文算法在两个数据集上的平均准确率

Table 2 Average accuracies of the algorithm for the two datasets

数据集	准确率/%			
	Split1	Split2	Split3	平均
UCF101	88.8	90.7	89.3	89.6
HMDB-51	62.3	60.9	61.9	61.7

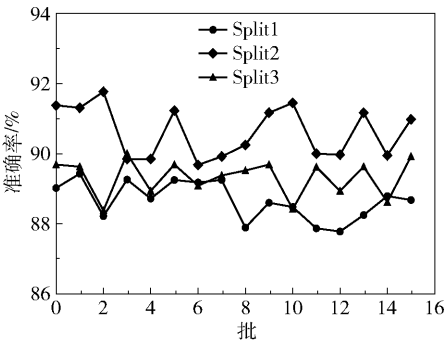


图 7 UCF101 测试集准确率曲线

Fig.7 Dataset accuracy curves for UCF101

表 3 为与经典的几种行为识别算法以及以不同数据形态为输入的算法的准确率对比,可以看出所设计的 Inflated VGGNet-16 网络在识别精度上均比经典的 iDT、two-stream 及 C3D 网络高,能够达到较高的识别准确率。利用 Kinetics 数据集对网络进行预训练以及双流卷积神经网络与 3D 卷积等不同的网络结构结合,都可以提高在 UCF101 等小型数据集上的识别准确率。也可以看出相比于 two-stream

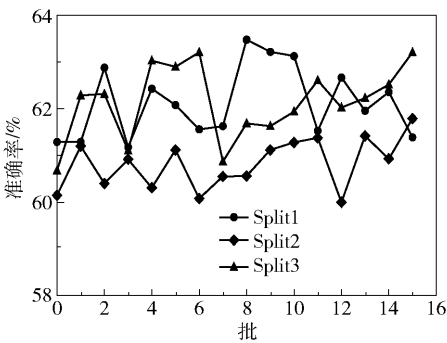


图 8  HMDB-51 测试集准确率曲线

Fig. 8  Dataset accuracy curves for HMDB-51

表 3  本文算法与当前先进算法对比

Table 3  Comparison with state-of-the-art methods

方法	准确率/%	
	UCF101	HMDB-51
iDT + FV <sup>[4]</sup>	85. 9	57. 2
two-stream <sup>[5]</sup>	88. 0	59. 4
Motion Vector + FV <sup>[15]</sup>	78. 5	46. 7
RGB + Enhanced Motion Vector <sup>[15]</sup>	86. 4	
RGB + RGB Diff <sup>[16]</sup>	87. 3	
two-stream I3D(Kinetics) <sup>[17]</sup>	98. 0	80. 9
C3D + liner SVM <sup>[18]</sup>	82. 3	
R(2 + 1)D-RGB(Kinetics) <sup>[19]</sup>	96. 8	74. 5
T-C3D(Kinetics) <sup>[20]</sup>	92. 5	62. 4
MARS + RGB + Flow(Kinetics) <sup>[26]</sup>	95. 8	
Inflated VGGNet-16	89. 6	61. 7

算法,所设计网络在 HMDB-51 数据集上的提升比较大,说明三维网络结构相比于二维网络在时间维度上的建模以及对运动信息的提取更加有利。

图 9 为用所设计网络对视频进行实时预测,左上角两行文字分别为预测动作类别及其相应的预测概率。从图中可以看出在背景变化小或者运动目标明显的动作中所设计网络具有非常高的识别准确度,在背景变化大以及运动目标较多或者不明显的情况下仍然具有较高的识别精度。

对于人体动作识别,类内差异性以及类间相似性问题的解决将会进一步提高识别精确度,利用分割的网络<sup>[31]</sup>可以将运动的突出部分进行区分并结合注意力机制的方法对视频中发生的行为进行识别。

5  结束语

本文设计的 Inflated VGGNet-16 网络可以有效

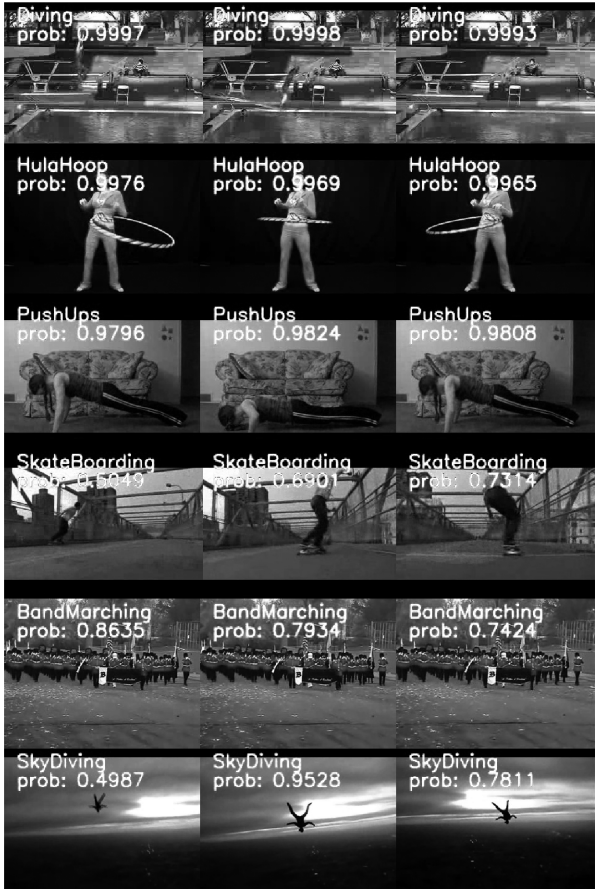


图 9  UCF101 数据集上算法识别效果

Fig. 9  Action recognition results for UCF101

利用 ImageNet 预训练模型,并结合批归一化层与 10 倍数据增强,加快网络收敛速度且降低了数据不足对网络性能的影响。该方法直接以连续的视频帧为网络输入,无需计算帧间光流,大大提高了识别速度。识别精度方面,所设计网络在 UCF101 与 HMDB-51 数据集上分别达到了 89.6% 和 61.7% 的准确率,比传统的 iDT 算法分别高 3.7% 与 4.5%,比 C3D 网络在 UCF101 上高 7.3%。

对于 3D 网络,即使对数据进行 10 倍增强,数据量还是不够,虽然可以通过增大视频数据集的规模来解决,但大型数据集对计算机资源的要求也会更高。因此寻找一种既能降低对计算机硬件要求又不损失分类性能的网络以及更加丰富有效的数据增强手段将是未来的研究方向。

参考文献:

[1] 孟勃,刘雪君,王晓霖. 基于四元数时空卷积神经网络的人体行为识别[J]. 仪器仪表学报, 2017, 38 (11): 2643 – 2650.  
MENG B, LIU X J, WANG X L. Human body action



- recognition based on quaternion spatial-temporal convolutional neural network [J]. Chinese Journal of Scientific Instrument, 2017, 38(11): 2643–2650. (in Chinese)
- [2] 罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 169–180.
- LUO H L, WANG C J, LU F. Survey of video behavior recognition [J]. Journal on Communications, 2018, 39(6): 169–180. (in Chinese)
- [3] WANG H, KLÄSER A, SCHMID C, et al. Dense trajectories and motion boundary descriptors for action recognition [J]. International Journal of Computer Vision, 2013, 103(1): 60–79.
- [4] WANG H, SCHMID C. Action recognition with improved trajectories [C] // Proceedings of the IEEE International Conference on Computer Vision. Sydney, 2013: 3551–3558.
- [5] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [C] // Advances in Neural Information Processing Systems. Montreal, 2014: 568–576.
- [6] 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. Las Vegas, 2016: 1933–1941.
- [7] DONAHUE J, HENDRICKS A L, 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. Boston, 2015: 2625–2634.
- [8] 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. Santiago, 2015: 4489–4497.
- [9] 李贤阳, 王建中, 杨竣辉, 等. 深度运动图耦合正则化表示的行为识别算法[J]. 电子测量与仪器学报, 2018, 32(1): 119–128.
- LI X Y, YANG J Z, YANG J H, et al. Action recognition algorithm based on depth motion maps and regularized representation [J]. Journal of Electronic Measurement and Instrumentation, 2018, 32(1): 119–128. (in Chinese)
- [10] 司阳, 肖秦琨, 李兴. 基于自动编码器和神经网络的人体运动识别[J]. 国外电子测量技术, 2018, 37(1): 78–84.
- SI Y, XIAO Q K, LI X. Human action recognition using auto-encoder and neural network [J]. Foreign Electronic Measurement Technology, 2018, 37(1): 78–84. (in Chinese)
- [11] 刘嘉莹, 张孙杰. 融合视频时空域运动信息的 3D CNN 人体行为识别[J]. 电子测量技术, 2018, 41(7): 43–49.
- LIU J Y, ZHANG S J. 3D CNN fusing spatial-temporal motion information in video for human action recognition [J]. Electronic Measurement Technology, 2018, 41(7): 43–49. (in Chinese)
- [12] 李庆辉, 李艾华, 王涛, 等. 结合有序光流图和双流卷积网络的行为识别[J]. 光学学报, 2018, 38(6): 0615002.
- LI Q H, LI A H, WANG T, et al. Double-stream convolutional networks with sequential optical flow image for action recognition [J]. Acta Optica Sinica, 2018, 38(6): 0615002. (in Chinese)
- [13] BILEN H, FERNANDO B, GAVVES E, et al. Dynamic image networks for action recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 3034–3042.
- [14] BILEN H, FERNANDO B, GAVVES E, et al. Action recognition with dynamic image networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2799–2813.
- [15] ZHANG B W, WANG L M, WANG Z, et al. Real-time action recognition with enhanced motion vector CNNs [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 2718–2726.
- [16] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition [C] // European Conference on Computer Vision. Cham: Springer, 2016: 20–36.
- [17] CARREIRA J, ZISSERMAN A. Quo Vadis, action recognition? a new model and the kinetics dataset [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, 2017: 6299–6308.
- [18] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks [C] // Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017: 5533–5541.
- [19] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 6450–6459.
- [20] LIU K, LIU W, GAN C, et al. T-C3D: temporal convolutional 3D network for real-time action recognition [C] // The

- Thirty-second AAAI Conference on Artificial Intelligence. New Orleans, 2018.
- [21] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild [J/OL]. (2012-12-03) [2019-11-01]. <https://arxiv.org/abs/1212.0402v1>.
- [22] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition [C]//2011 International Conference on Computer Vision. Barcelona: IEEE, 2011: 2556–2563.
- [23] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 6546–6555.
- [24] KÖPÜKLÜ O, KOSE N, GUNDUZ A, et al. Resource efficient 3D convolutional neural networks [J/OL]. (2019-09-09) [2019-11-01]. <https://arxiv.org/abs/1904.02422>.
- [25] TRAN D, WANG H, TORRESANI L, et al. Video classification with channel-separated convolutional networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019: 5552–5561.
- [26] CRASTO N, WEINZAEPEL P, ALAHARI K, et al. MARS: motion-augmented RGB stream for action recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, 2019: 7882–7891.
- [27] HEILBRON F C, ESCORCIA V, GHANEM B, et al. Activitynet: a large-scale video benchmark for human activity understanding [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 961–970.
- [28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J/OL]. (2015-04-10) [2019-11-01]. <https://arxiv.org/abs/1409.1556>.
- [29] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [J/OL]. (2015-03-02) [2019-11-01]. <https://arxiv.org/abs/1502.03167>.
- [30] TU Z G, XIE W, QIN Q Q, et al. Multi-stream CNN: learning representations based on human-related regions for action recognition [J]. Pattern Recognition, 2018, 79: 32–43.
- [31] LI R R, LIU W J, YANG L, et al. DeepUNet: a deep fully convolutional network for pixel-level sea-land segmentation [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11 (11): 3954–3962.

## Inflated VGGNet-16 networks for human action recognition

WANG Zhen LIU RuiMin\* HUANG QiongTao

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China)

**Abstract:** Human behavior recognition 3D networks suffer from problems of shallow C3D network structure, poor feature extraction ability, lack of available pre-training models, and long training times. By starting from a deeper VGGNet-16 network, and adding a batch normalization layer and using the ImageNet pre-training model Inflating method for network initialization, we have designed a new human behavior recognition 3D network. In experimental analysis using the standard datasets UCF101 and HMDB-51, images were center-cropped and used as the input to the network. The accuracy of the original C3D network was 9.2% higher using the UCF101 dataset from scratch, and the network convergence was faster, which shows that our Inflated VGGNet-16 network has stronger feature extraction and better generalization capabilities. Finally, our network was modified with ten-fold data enhancement, and the accuracy ratio for the two standard data sets UCF101 and HMDB-51 was 89.6% and 61.7% respectively, which in the case of UCF101 is 7.3% higher than the value for the shallower C3D network, and have higher behavior recognition accuracy than the traditional improved dense trajectory method (iDT) and the classic two-stream convolutional neural network.

**Key words:** action recognition; VGGNet-16; Inflating; ImageNet pre-training; data augmentation