



综述

基于深度学习的视频行为识别方法综述

赵朵朵¹, 章坚武¹, 郭春生¹, 周迪², 穆罕默德·阿卜杜·沙拉夫·哈基米¹

(1. 杭州电子科技大学, 浙江 杭州 310018; 2. 浙江宇视科技有限公司, 浙江 杭州 310018)

摘要: 近年来, 自动学习特征的深度学习方法在视频行为识别领域中不断被挖掘探索。在总结了常用的行为识别数据集的基础上, 对传统的行为识别方法以及深度学习的相关基础原理进行了概述, 着重对基于不同输入内容与不同深度网络的行为识别方法进行了较为全面、系统性的总结、对比与分析。最后, 对深度学习在行为识别领域的发展做了总结并展望了未来的发展趋势。

关键词: 行为识别; 数据集; 自动学习; 深度网络

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019286

A survey of video behavior recognition based on deep learning

ZHAO Duoduo¹, ZHANG Jianwu¹, GUO Chunsheng¹, ZHOU Di²,

MOHAMMED ABDU SHARAF ALHAKIMI¹

1. Hangzhou Dianzi University, Hangzhou 310018, China

2. Zhejiang Uniview Technologies Co., Ltd., Hangzhou 310018, China

Abstract: In recent years, the deep learning method of automatic learning features has been continuously explored in the field of video behavior recognition. The traditional behavior recognition methods and the underlying principles of deep learning were outlined. Then a number of behavior recognition methods based on different input content and different deep networks was compared and analyzed. Finally, the development of deep learning in the field of behavior recognition was concluded and its future development trend was prospected.

Key words: behavior recognition, dataset, automatic learning, depth network

1 引言

以各种互动的方式产生的行为是人最基本的

活动, 视频行为识别是一种“读懂行为”, 即读懂视频中的内容并做出一定的判断, 是视频理解的一部分。其最终目的是分析视频中有什么人、

收稿日期: 2019-08-19; 修回日期: 2019-12-10

通信作者: 章坚武, jwzhang@hdu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61772162, No.U1866209); 国家重点研发计划资助项目 (No.2018YFC0831503); 浙江省自然科学基金资助项目 (No.LY16F020016); 浙江省重点研发计划资助项目 (No.2018C01059, No.2019C01062)

Foundation Items: The National Natural Science Foundation of China (No.61772162, No.U1866209), The National Key Research Development Program of China (No.2018YFC0831503), The Natural Science Foundation of Zhejiang Province of China (No.LY16F020016), The Key Research Development Program of Zhejiang Province of China (No.2018C01059, No.2019C01062)



在什么时间、什么地方做了什么事情。视频行为识别的难点主要在于：（1）行为持续时间的不确定性以及行为的多速率问题，尤其是速度变化大、行为的起始时间难以确定；（2）行为的多样性，一个人的动作多样，同一个动作不同的人表现不同；（3）背景的复杂性，视频图像视角差异大，分辨率参差不齐，存在遮挡抖动等问题。正是这些难点将专家学者们的注意力从传统的行为识别方法转移到深度网络中。

先构建网络架构学习数据的各层次特征的抽象表征，再通过反向传播（back propagation, BP）算法对构建的模型进行训练的过程叫做深度学习。深度学习作为人工智能（AI）的热门研究领域，能够通过监督学习与弱监督学习自动学习视

频中的动作特征。将其应用于视频行为处理领域可以增强网络表征能力和泛化能力，开展更为高效和精准的视频行为识别工作。

2 相关基础

2.1 常用的数据集介绍

近年来，为了评估行为识别方法的性能，提升深度学习在视频行为识别中的应用价值，研究者们创建了大量的公开数据集。这对促进人体行为识别方法的研究起到了关键性的推动作用。目前行为识别常用的数据集可分为通用数据集和真实场景数据集两大类，各种常用数据集的详细介绍见表1（2004–2018年）。其中KTH^[1]和Weizmann^[2-3]数据集属于通用行为识别数据集，是

表1 数据集汇总

名称（年份）	简介	视频样本数	目前最高识别率
KTH ^[1] （2004年）	由25人完成的6类动作：挥手、步行、慢跑、跑步、拍掌和拳击。有4个不同场景	2 391	98.83% ^[14]
Weizmann ^[2-3] （2005年）	包含9人完成的10个动作：走、跑、跳、挥手和弯腰等	93	100% ^[15-16]
Hollywood 2 ^[4] （2009年）	含10类场景下的12种行为类别：打架、开车、握手、拥抱、亲吻等。样本均来自69部Hollywood电影	3 669	78.6% ^[17]
Olympic sports ^[5] （2010年）	含16种运动动作，有跳高、跳远等	783	96.6% ^[18]
HMDB51 ^[6] （2011年）	含51个类别，一个类别至少含101段视频样本。样本来源于公共数据库、电影和YouTube等	6 849	82.1% ^[19]
UCF Datasets （2007年–）	（1）UCF-11(2008) ^[7] （2008年） 有11种行为类别：篮球投篮、与狗一起散步等	1 600	94.5% ^[20]
	（2）UCF-50 ^[8] （2009年） 类别由11种扩展至50种。视频来自YouTube，是UCF-11的扩展	6 676	99.98% ^[21]
	（3）UCF-101 ^[9] （2012年） 是UCF-50的扩展，类别由50种扩展至101种。分为五大类：人与人交互、人与物交互、弹奏乐器、运动及肢体运动	13 320	98.0% ^[22]
	（4）UCF Sports ^[10] （2013年） 含10种运动的类别，有踢足球、举重、跳水等。收集于广播电视频道以及互联网等的各类运动样本	150	96.2% ^[23]
Sports-1M ^[11] （2014年）	包含487种体育运动项目，分为六大类：水上、团队、冬季、球类、对抗、与动物等运动	1 133 158	75.9% ^[24]
Kinetics ^[12] （2017年）	覆盖700种人类动作，每个动作至少包含600个视频，每一段都来自一个独特的YouTube视频。动作包括演奏乐、握手、拥抱等	650 000	Top-1:83.5%, Top-5: 96.8% ^[25]
Moments in Time ^[13] （2018年）	涉及人物、动物、物体或自然现象，捕捉了动态场景中标记了的3秒视频	1 000 000	Top-1:38.1%, Top-5:65.3% ^[26]

注：Top-1指预测结果中只有概率第一的分类正确才表示预测正确，否则预测错误；Top-5指预测结果中概率的前5中只要有正确的分类就表示预测正确，否则预测错误

受试者在受限的场景下做出的一系列动作。这类数据集背景单一、动作种类少,没有包含相机运动,与真实的场景差别很大。真实场景数据集收集于电影、YouTube 等视频中,由于相机移动、场景不固定且动作类内离散度大而极具挑战性。它包括 Hollywood^[4]、Olympic sports^[5]、HMDB51^[6]、UCF^[7-10]、Sports-1M^[11]、Kinetics^[12]、Moments in Time^[13]等。从数据集的发展趋势来看,近几年数据集里包含越来越多的动作类别与样本数目,纷繁复杂的视频场景,使其越来越接近于不受控的自然状态下的情景。这种更接近实际情况的发展趋势使得视频行为识别更具实用价值,但同时对算法的稳健性也提出了更高的要求。目前行为识别领域针对 UCF-101 和 HMDB51 数据集的研究有很多并且相当深入,因此,本文主要是对基于这两个公开数据集的方法做了比较全面的阐述、对比和分析^[14-26]。

2.2 传统行为识别技术概述

传统的视频分类方法主要是基于手工提取特征的识别方法,处理过程一般包括特征采样、选取描述符、特征预(后)处理、描述符聚类 and 向量编码等步骤^[27]。它的处理流程如图 1 所示。在深度学习进入行为识别领域之前,由 Wang 等^[28-29]提出的改进的密集轨迹(improved dense trajectories, IDT)算法是效果较好、经典的传统算法之一。它将数据集 UCF-50 的准确率从 84.5%提高到了 91.2%,将 HMDB51 提高了 0.6 个百分点。

IDT 算法是通过每一帧的光流信息得到视频中物体的运动轨迹,再沿着运动轨迹提取轨迹变

化(trjectories)、方向梯度直方图(histogram of oriented gradient, HOG)、运动边界直方图(motion of boundary history, MBH)和光流梯度直方图(histogram of oriented optical flow, HOF)4 种特征;然后对连续帧中的每一个像素点计算场景中所有物体的三维运动向量在所在平面上的二维投影的向量,它包含了物体运动的速度和方向;接着从初始帧开始,每隔一定数量的像素进行采样跟踪,由光流判断跟踪点在下一帧的位置;最后采用费舍尔编码(fisher vector, FV)对特征进行编码,再基于编码结果训练 SVM 分类器。

IDT 算法实现简单、分类效果良好,稳定性强、可靠性高,缺点是其特征维度较高,对运算资源消耗过大。尽管该方法对行为的表示能力远不如深度网络,但是它沿着轨迹提取特征、通过估计相机运动来减弱其带来影响的思路至今仍非常值得借鉴。

2.3 神经网络概述

2.3.1 卷积神经网络概述

卷积神经网络(convolutional neural network, CNN)是带有卷积结构的深度神经网络,它由输入、卷积、激活函数、池化和全连接 5 个层组成。其中,输入层主要是对原图像进行白化、归一化、去均值等预处理操作;卷积运算层主要使用滤波器(卷积核)进行随机初始化,通过设定步长、深度等参数,对输入进行不同层次的局部特征提取;激活函数主要是把卷积层的输出做非线性映射。常见的激活函数有 ReLU、Sigmoid、Tanh 等;池化层主要用于特征降维,通过压缩数据与参数

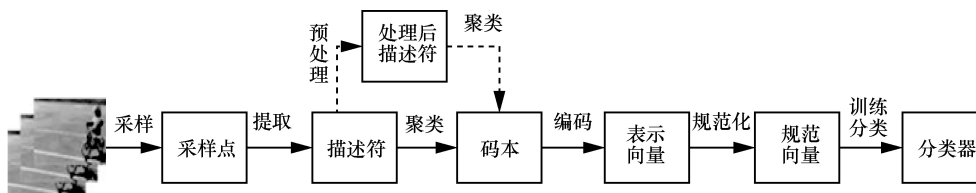


图 1 传统行为识别流程



数量来减小过拟合, 提高模型的容错性。常见池化方式有平均池化 (mean pooling) 和最大值池化 (max pooling); 全连接层将每个结点与上层的所有结点相连, 综合前面提取到的特征来解决非线性问题, 实现分类。CNNs 中的经典网络有 LeNet-5^[30]、AlexNet^[31]、VGG-16^[32]、ResNet^[33]、GoogLeNet^[34]等。

二维卷积神经网络 (2D CNN) 是将每帧图像通过 CNN 来识别, 由于忽略了时间维度上的帧间运动信息而无法对时间信息和运动模式建模。然而, 运动信息在视频行为识别中有着关键性的作用^[35]。为了有效地综合运动信息, Ji 等^[36]提出 3D 卷积 (3D Conv) 方法来捕获视频中的时间特征信息。该结构卷积层中的每个特征图谱 (feature map) 都与上层中多个邻近的连续帧相连, 通过增加时间维度赋予神经网络行为识别的功能。2D 卷积与 3D 卷积的对比如图 2 所示。

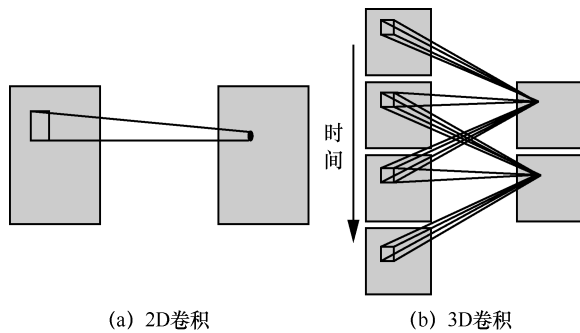


图 2 2D 与 3D 卷积对比

2.3.2 循环神经网络概述

视频处理中输入的训练样本一般是基于时间、长短不一的连续序列数据, 直接的拆分成一个个独立的样本来通过 CNN 进行训练比较困难。与基础的只在层与层之间建立全连接的网络不同, 循环神经网络 (recurrent neural network, RNN) 通过在层之间的神经元之间建立连接能够对未分割的序列定位行为动作的起止点和判定动作的类型。RNN 的结构如图 3 所示, 它是一个单元结构的复用, 每个圆圈表示一个单元,

每个单元的功能一致。其中一般将输入单元的输入集标记为 $\{x_0, x_1, \dots, x_t, x_{t+1}, \dots\}$, 将隐藏单元的输出标记为 $\{s_0, s_1, \dots, s_t, s_{t+1}, \dots\}$, t 时刻的记忆 s_t 可根据当前输入层的输出与上一步隐藏层的状态进行计算如式 (1):

$$s_t = f(U \times x_t + W \times s_{t-1}) \quad (1)$$

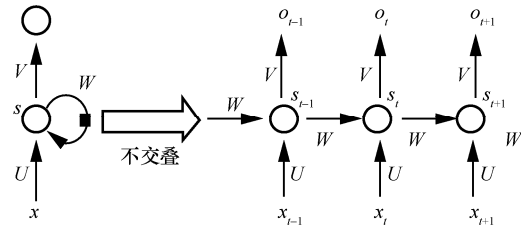


图 3 循环神经网络结构

输出单元的输入标记为 $\{o_0, o_1, \dots, o_t, o_{t+1}, \dots\}$, t 时刻的输出 o_t 计算如下:

$$o_t = \text{Softmax}(Vs_t) \quad (2)$$

其中, f 一般是非线性的激活函数。 U 、 W 、 V 是权重矩阵参数, 每输入一步输入中各层都共享这些参数。一般 s_t 只需包含前面若干步的隐藏层状态。

在 RNN 中, 由 Hochreiter^[37]提出的长短期记忆 (long short-term memory, LSTM) 模型是目前比较常用的, 该模型能够对长短时依赖进行更好的表达, 在视频识别中可以用来为每一帧添加标签和判断类别, 解决了 RNN 中梯度消亡的问题。

3 基于深度学习的视频行为识别方法

3.1 基本思想概述

近年来, 深度学习方法在视频行为识别领域应用广泛, 影响该方法性能的因素也有很多, 其中输入端输入的内容、网络的结构是影响其性能的主导因素, 基于深度学习方法的基本流程如图 4 所示。视频中一般都含有彩色 (RGB) 图像信息、运动光流 (optical) 信息以及声音信息 (audio), 不同的信息在不同的视频行为识别中所占的权重不同, 本文把这些信息统称为输入内容。网络结

构是基于输入内容构建的不同深度的神经网络以及网络中层与层之间不同的交互方式。深度学习发展的热潮催生了很多应用于不同任务的优秀网络结构,在视频行为分类领域,基于行为的时空属性产生了 3D 卷积网络(3D convolutional network, C3D)和双流卷积网络(two-stream convolutional network)两个主流的网络架构。下面将对基于不同输入、不同网络的行为识别方法进行阐述对比,尤其对主流网络及其延伸网络的各种视频行为识别方法进行分析。

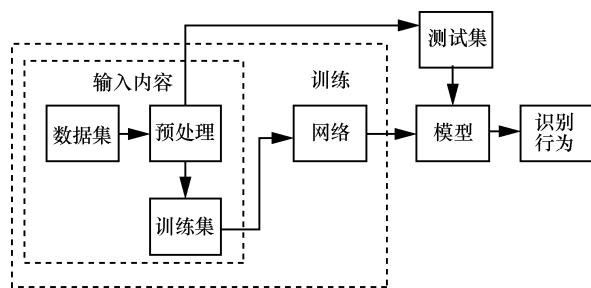


图4 基于深度学习的视频行为识别流程

3.2 基于不同输入内容的行为识别方法

随着目标检测技术发展越来越成熟,可以从单帧或堆叠的多帧 RGB 图像中获得一部分视频集上的行为类别^[38]。如参考文献[39]提出的在 KTH 数据集上自动学习特征的深度模型以及参考文献[40]提出的基于 LSTM 的递归神经网络模型,均是把彩色序列图像放入 CNN 网络中进行训练,最后经过 Softmax 层输出视频分类结果。但是,一方面由于同一个视频片段背景一般比较单一,相邻多帧的差异性很小,使得训练集数据的多样性差、冗余量大;另一方面仅仅输入 RGB 图像信息没有充分利用整个视频的上下文信息而存在很多的漏检目标。而光流信息能够将当前帧的检测结果向前向后传的特点,对视频行为分类来说是一个很好的特征。目前使用神经网络获取光流的方法很流行,比如 Ilg 等^[41]提出的光流神经网络 FlowNet2.0。它是 FlowNet^[42]的增强版,它整体上速度代价变小,性能得到了大幅度提升,追平了

目前领先的传统方法。把 RGB 和运动光流图像信息放入 CNN 网络中,经过数据融合后输出结果,融合的点不同,融合的方式等存在不同。参考文献[37,43-44]等用不同的深度卷积神经网络、不同的网络融合方式分析了彩色图像序列和光流图像的融合。

音频信息作为视频的关键部分也对视频中的行为识别起着关键性的作用。处理音频特征一般采用基于卷积网络的音频分类系统:先把每个视频分成若干帧,再通过傅里叶变换、直方图积分和对数变换等提取其频域信息,最后将每个视频的语音信息以一定的形式输入分类网络,生成标签概率分布预测。由于语音信息难以增强,与训练集容易过度拟合,一般易采用复杂度较低的网络提取。近年来,通过对音频信息分析为其添加语义标签后加入深度网络自动学习音频特征的方式,能够在一定程度上提高识别的准确率。

基于 RGB 图像信息和运动光流信息以及音频信息的识别方式叫做多模态识别,基于多模态识别网络探索方式也有很多,其中网络结构不同,全连接层融合后的输出不同。参考文献[45-47]等均在双流网络的基础上加入声音信息,对这三流信息分别建模探索了新的融合网络架构并取得了不错的成绩。在 CVPR 2018 视频行为识别挑战赛上海康威视^[26]、旷视科技^[25]以及七牛云团队均运用多模态识别方式分别斩获前三名,尤其是海康威视提出的建立在空间网络和 3D 卷积神经网络基础上的 DEEP-HR 网络,输入端使用多模态输入方式获得了比赛的第一名。

3.3 基于不同网络的行为识别方法

3.3.1 双流网络及其衍生方法

CNNs 在视频处理领域上有很强的学习和表现视频特征的能力,但是随着网络规模的扩大、参数的增加、大量标签数据的产生,使得 CNN 在大规模视频分类上的性能受到挑战。针对视频行为难分类的问题,Simonyan 等^[48]提出了基于时间



和空间维度的双流 CNN。双流 CNN 对视频信息的理解是通过效仿人体视觉过程进行的, 它以单个包含人物与环境的视频图像帧作为表述空间信息的载体组成空间信息网络, 提取形状等特征信息; 光流信息作为时序信息的载体输入到另外一个 CNN 中, 称为时间信息网络, 用来处理连续多帧密集光流, 提取动作信息。后续采用多任务训练的方法处理两个流, 经过 Softmax 后做分类得分融合。

对于双流网络的改进, Ng 等^[37]提出采用 LSTM 来聚合时间轴上帧的 CNN 最后一层的激活, 如图 5 所示。该方法通过 LSTM 引入的记忆单元能够更有效地表达帧的前后顺序。大多数研究选择在 CNN 全连接层的前一层进行特征融合, 因为如果取其后的多层特征进行池化, 空间特征在时间轴上的信息会丢失。相对参考文献[48], 它加长了 CNN 特征融合的时间, 能对更长时长的视频进行表达。Feichtenhofer 等^[49]提出将时空网络融合在一起, 如图 6 所示, 将双流通过一个过滤器结合在一起, 该过滤器能够从时间流中学习高度抽象的空间流外观特征和短期运动特征之间的对应关系, 然后多个长

期输入块在最后一个卷积层融合。在最后一个卷积层, 通过 3D 卷积融合 (3D conv fusion) 和 3D 池化 (3D pooling) 将其转换为时空流, 同时不截断时间流, 在时间网络中执行 3D pooling。最后所得到的时空特征在时空中被三维汇聚再传递给全连接层分别计算分类损失。

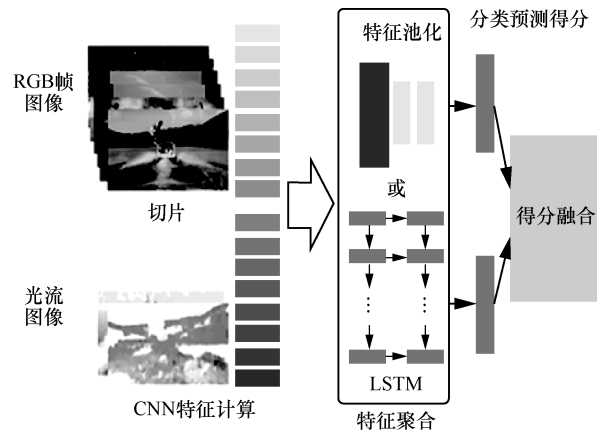


图5 结合 LSTM 的双流改进网络

以上基于双流网络的架构都需要密集采样视频帧, 庞大的计算量一直是难以逾越的鸿沟。对此, Wang 等^[50]利用稀疏时间采样, 提出了时间段网络 (temporal segment networks, TSN)。该网络

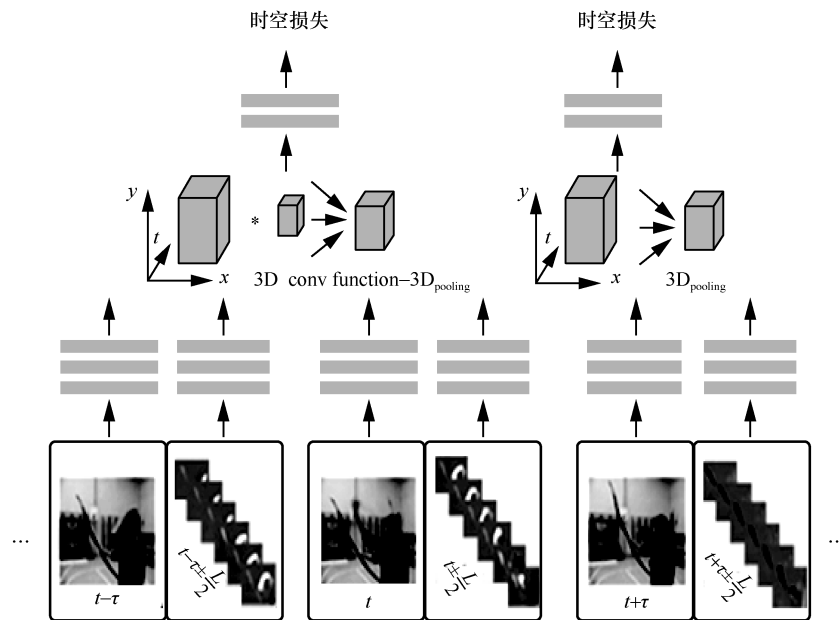


图6 双流融合网络结构

继承了双流网络的结构,通过使用多个双流网络,分别捕捉不同时序位置的短时段信息来处理长时段信息表达的问题。TSN 网络结构如图 7 所示,它在输入端不考虑视频的时长,直接将视频切割为 K 个段,随机地在每个段中找出一个小的时间片;然后用 CNN 这些时间片分别提取时空特征;而后在时间和空间上分别进行段共识的特征级融合,最后再进行 Softmax 分类。TSN 计算原理如下:

$$\text{TSN}(X_1, X_2, \dots, X_K) = S(R(f(X_1; W), F(X_2; W), \dots, F(X_K; W))) \quad (3)$$

其中, X_K 代表第 K 个段, W 为参数, R 是特征融合函数,函数 F 指经过 CNN 提取的特征, S 表示分类层 Softmax。

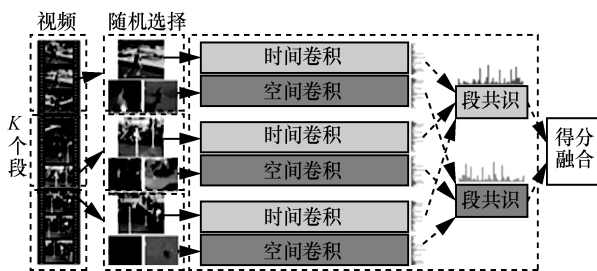


图 7 TSN 结构

TSN 网络大大降低了双流网络计算的复杂度,研究者们对其进行了进一步的研究探索。Lan 等^[51]是对 TSN 的融合部分做了改进,提出由深度网络自学习决定的加权融合方法,该方法认为片段不同所占的权重不同,且权重的大小由网络学习获得。而 Zhou 等^[52]则关注时序关系的推理,提出通过视频帧之间的时序推理可以识别仅靠关键帧无法辨别的动作。Diba 等^[53]结合前面两者的改进措施,增加三层全连接层学习不同视频片段的权重,在输入的特征图上做时序推理,并提出一种时序线性编码层 (temporal linear encoding, TLE) 来对视频分段提取后的特征图进行融合编码,对所有的帧进行综合编码获得视频表达,捕捉到长时间动态过程。

3.3.2 C3D 卷积网络及其衍生方法

对于时空特征的学习,Tran 等^[54]在大规模、有监督数据集上提出了一种简单有效的方法——C3D。它将时间看成是第三维信息来融合单帧的 RGB 图像和帧与帧之间的变化信息。它的网络结构非常简单,含卷积层 8 个、池化层 (max pooling) 5 个、全连接层 2 个。尽管 3D 卷积神经网络通过简单的线性分类器就能达到不错的性能,但较大的内存消耗和较高的计算成本导致其难以构建更深层的 3D CNN。

针对 3D 网络的局限性,许多研究学者对其进行了探索。最早由 Sun 等^[55]从卷积网络的卷积层入手,将 3D 空间时间学习分解为 2D 空间和 1D 时间学习,提出了一种时空分解卷积网络 (factorized spatio-temporal convolutional network, $F_{st}CN$)。该网络级联地组合视频中的时空信息。在网络的底层,使用 2D 空间卷积核学习视频帧的空间表象特征。在网络的顶层,使用 1D 时间卷积核学习视频的时间运动特征。这种网络结构与 3D 卷积网络相比大大减少了参数量,所以对训练数据量要求比较低。从卷积核结构角度出发, Qiu 等^[56]对 ResNet 网络内部连接的卷积形式做了改进,构建了 P3D (pseudo-3D) 模型,经济有效地表达了视频特征; Diba 等^[57]则引入一种新的时域层 (temporal transition layer, TTL) 给可变时域卷积核深度建模,提出 T3D (temporal 3D) 模型,用不同尺度的卷积来捕捉信息并采用迁移学习节省训练时间。Carreira 等^[58]提出了一种基于 GoogLeNet 的 Inception-V1 网络的 I3D (two-stream inflated 3D convNets) 模型,在 UCF-101 和 HMDB51 数据集上分别获得了 98% 和 80.9% 的识别率。Tran 等^[59]沿用参考文献[54-55]的想法将时空卷积分开,分别得到时间层和空间层,提出了新的时空卷积块 $R(2+1)D$ 。不同的是它在网络所有层中使用的都是统一类型的时空残差块,该模型在识别精度上略差于 I3D。



3.3.3 其他优秀网络

针对目前对光流特征依赖性较强且对于时空数据的表征学习相对困难的问题, Fan 等^[60]创造性地提出了一种从数据中学习类光流特征、基于端到端训练的神经网络 TVNet。该网络性能高效,不但能自然地连接嫁接到一些特定任务网络,而且无需预计算和预存储。目前,研究者们对视频行为识别中存在的多速率问题也做了很多研究,尤其是 Zhu 等^[19]提出了时间金字塔池深度网络(deep networks with temporal pyramid pooling, DTPP),一种端到端的视频级表示学习方法。它的网络结构如图 8 所示,其中 TPP 层是把视频的帧级特征整合到固定大小的视频级表示中,以多尺度的方式获取视频的时间结构。该模型的视频级表示紧凑,具有多个时间尺度,可全局感知。目前, DTPP 在 UCF-101 和 HMDB51 数据集上,无论是通过 ImageNet 还是 Kinetics 预训练,都取得了很好的效果。尽管 DTPP 取得了最优的性能。但以它为代表的部分深度网络都是学习一整段的动作视频,由于一些帧与动作的关系并不大,大量不相关内容累积的计算量严重影响了训练的速度。对此, Zhu 等^[61-62]提出了关键卷(key volume)特征的自动识别。它通过在关键帧上建立 CNN 模型,增强了动作分类的效果。而 Kar 等^[63-64]则以在视频中除去一些冗余帧,提高视频行为的判断能力为原则,提出了一种关键视频帧汇聚方法。近年来,这种基于注意力机制提取视频关键特征方法引起了广泛的关注。

4 识别方法分析评价

本文主要在 UCF101 和 HMDB51 数据集上来评价视频行为识别方法的性能,根据第 3.3 节中的网络分类方法,对双流及其衍生网络行为识别方法、C3D 卷积网络及其衍生行为识别方法以及其他优秀网络行为识别方法在 UCF-101 和 HMDB51 数据集上的性能进行了比较和汇总,见表 2、表 3

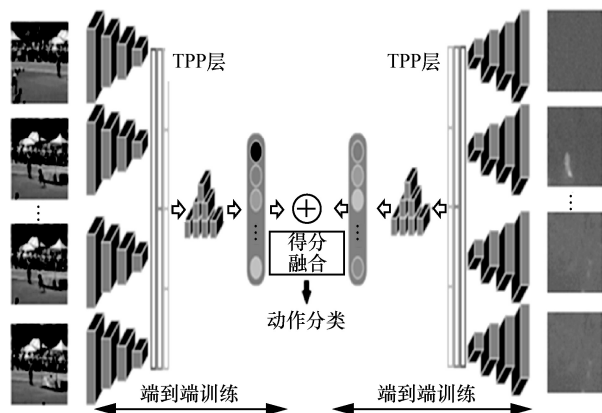


图 8 DTPP 网络结构

和表 4。由汇总结果可见,尽管早些年基于深度学习的视频行为识别方法在识别准确率上不如传统方法,但是近些年来有了突破性进展:UCF101 数据集上的准确率达到 98%,比传统最好的方法高出 12.1 个百分点;在 HMDB51 数据集上达到 80.9%,超过 IDT 算法 23.7%。另外,由表 2、表 3 和表 4 可见,深度网络与 IDT 算法相结合能够提升网络的性能,使用的预训练数据集的不同也会影响网络的识别准确率。

5 结束语

本文基于深度学习的视频行为识别方法从不同的输入内容和不同的深度网络两个方面出发,通过对目前的主流网络以及它们的延伸网络进行了对比分析,发现影响检测性能的因素有很多——输入网络的信息、输入信息的融合方式、不同网络之间的相互融合、不同网络层次之间的交互以及基于不同数据集的预训练都会影响识别的准确率。值得注意的是,输入端在考虑图像信息、运动光流信息以及音频信息的多模式融合的同时应该注意到不同的信息在不同视频行为识别中所占的权重不同。另外,光流信息的替代问题也值得进一步地探究。

目前基于深度学习的视频行为识别方法已经取得了一定的研究成果,已经有很多的优秀网络架构被研究公开,但是仍然面临一些挑战。随着

表2 双流网络及其衍生行为识别方法在UCF-101和HMDB51数据集上的性能比较

参考文献	方法	预训练数据集	UCF-101	HMDB51
[48]	Two-Stream(VGG-M)	ImageNet	88.0	59.4
[37]	Two-Stream+LSTM	ImageNet	88.6	—
[64]	Very deep Two-Stream(GoogLeNet)	ImageNet	89.3	—
	Very deep Two-Stream(VGG-16)		91.4	58.5
[49]	Two-Stream fusion (VGG-16)	ImageNet	92.5	65.4
	Two-Stream fusion (VGG-16) + IDT		93.5	69.2
[65]	ST-ResNet*	ImageNet	93.4	66.4
	ST-ResNet*+IDT		94.6	70.3
	ST-Pyramid Network(VGG-16)		93.2	66.1
[66]	ST-Pyramid Network(ResNet-50)	ImageNet	93.8	66.5
	ST-Pyramid Network(BN-Inception)		94.6	68.9
[67]	ST-Multiplier	ImageNet	94.2	68.9
	ST-Multiplier+IDT		94.9	72.2
[53]	TLE: FC-Pooling	ImageNet	92.2	68.8
	TLE: Bilinear+TS		95.1	70.6
	TLE: Bilinear		95.6	71.1
[50]	TSN	ImageNet	94.2	69.4
	TSN(Inception v3)		96.2	75.3
[68]	Hidden Two-Stream (TSN)	Kinetics	93.2	66.8
	Hidden Two-Stream (I3D)		97.1	78.7

表3 C3D 卷积网络及其衍生行为识别方法在UCF-101和HMDB51数据集上的性能比较

参考文献	方法	预训练数据集	UCF-101	HMDB51
[55]	F ₃ CN	ImageNet	88.1	59.1
[54]	C3D one network	Sports-1M	82.3	—
	C3D ensemble		85.2	—
	C3D ensemble + IDT		90.1	—
[35]	C3D+LSTM	—	92.9	70.1
[57]	T3D	ImageNet	90.3	59.2
	T3D-Transfer		91.7	61.1
	T3D+TSN		93.2	63.5
[38]	STRN	ImageNet	93.2	64.9
[56]	P3D ResNet	ImageNet+Sports-1M	88.6	
	P3D ResNet+IDT		93.7	
[68]	Multi-task C3D+LSTM	Sports-1M	93.4	68.9
[59]	R(2+1)D-RGB	Sports-1M	93.6	66.6
	R(2+1)D-Flow		93.3	70.1
	R(2+1)D-Two-Stream		95.0	72.7
	R(2+1)D-RGB	Kinetics	96.8	74.5
	R(2+1)D-Flow		95.5	76.4
	R(2+1)D-Two-Stream		97.3	78.7
[58]	RGB-I3D	ImageNet+Kinetics	95.6	74.8
	Flow-I3D		96.7	77.1
	Two-Stream I3D		98.0	80.7
	RGB-I3D	Kinetics	95.1	74.3
	Flow-I3D		96.5	77.3
	Two-Stream I3D		97.8	80.9



表 4 其他优秀网络行为识别方法在 UCF-101 和 HMDB51 数据集上的性能比较

参考文献	方法	预训练数据集	UCF-101	HMDB51
[69]	TDD	ImageNet	90.3	63.2
	TDD+IDT		91.5	65.9
[70]	LTC	Sports-1M	91.7	64.8
	LTC+IDT		92.7	67.2
[61]	Key -volume mining deep framework	ImageNet	93.1	63.3
[62]	AdaScan	ImageNet	89.4	54.9
	AdaScan+iDT		91.3	61.0
	AdaScan+ iDT+C3D		93.2	66.9
[71]	RNN-FV(C3D+VGG-CCA)	—	54.33	88.01
	RNN-FV(C3D+VGG-CCA) +IDT	—	94.08	67.71
[51]	DOVF	ImageNet	94.9	71.7
	DOVF + MIFS		95.3	75.0
[60]	TVNet	—	94.35	71.0
	TVNet+IDT		95.4	72.6
[72]	Four-Stream	ImageNet	95.5	72.5
	Four-Stream + IDT		96.0	74.9
[19]	DTPP	ImageNet	95.8	74.8
	DTPP + MIFS		96.1	76.3
	DTPP + IDT		96.2	75.3
	DTPP	Kinetics	98.0	82.1

数据集的不断扩充, 对于硬件系统的性能也提出了更高的要求, 应考虑如何在小样本数据中就能获得较好的识别准确度以及在大样本集中如何进一步提高速度等; 对于一些实际应用, 比如对于异常行为的识别、细微动作识别等这些更加贴近实际的应用还存在很大的困难, 而实际场景中得到应用必须考虑算法高效性, 如何建立更简单的网络模型, 训练优化更少的参数提升算法收敛的速度以及如何在视频中提取出描述视频判断的强有力的特征等都是下一步的重点研究方向。

参考文献:

- [1] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]//17th International Conference on Pattern Recognition(ICPR), Aug 23-26, 2004, Cambridge, UK. Piscataway: IEEE Press, 2004: 32-36.
- [2] BLANK M, GORELICK L, SHECHTMAN E, et al. Actions as space-time shapes[C]// 10th IEEE International Conference on Computer Vision(ICCV), Oct 17-21, 2005, Beijing, China. Piscataway: IEEE Press, 2005: 1395-1402.
- [3] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as space-time shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2247-2253.
- [4] MARSZALEK M, LAPTEV I, SCHMID C. Actions in context[C]//22nd IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jun 20-25, 2009, Florida, USA. Piscataway: IEEE Press, 2009: 2929-2936.
- [5] NIEBLES J C, CHEN C W, LI F F. Modeling temporal structure of decomposable motion segments for activity classification[C]// 11th European Conference on Computer Vision (ECCV), Sep 5-11, 2010, Heraklion, Crete, Greece. Berlin: Springer Verlag, 2010: 392-405.

- [6] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//16th IEEE International Conference on Computer Vision (ICCV), Nov 6-13, 2011, Barcelona, Spain. Piscataway: IEEE Press, 2011: 2556-2563.
- [7] LIU J G, LUO J B, SHAH M. Recognizing realistic actions from videos[C]//22nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 20-25, 2009, Florida, USA. Piscataway: IEEE Press, 2009: 1996-2003.
- [8] REDDY K K, SHAH M. Recognizing 50 human action categories of Web videos[J]. Machine Vision and Applications, 2013, 24(5): 971-981.
- [9] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. Computer Science, 2012: 1-7.
- [10] RODRIGUEZ M D, AHMED J, SHAH M. Action match a spatio-temporal maximum average correlation height filter for action recognition[C]//21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 24-26, 2008, Anchorage, Alaska, USA. Piscataway: IEEE Press, 2008: 1-8.
- [11] KARPATY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 23-28, 2014, Columbus, USA. Piscataway: IEEE Press, 2014: 1725-1732.
- [12] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. 2017. arXiv: 1705.06950.
- [13] MONFORT M, ANDONIAN A, ZHOU B, et al. Moments in time dataset: one million videos for event understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019(3): 1-9.
- [14] XU W R, MIAO Z J, TIAN Y. A novel mid-level distinctive feature learning for action recognition via diffusion map[J]. Neurocomputing, 2016(218): 185-196.
- [15] TONG M, WANG H Y, TIAN W J, et al. Action recognition new framework with robust 3D-TCCHOGAC and 3D-HOOFGAC[J]. Multimedia Tools and Applications, 2017, 76(2): 3011-3030.
- [16] VISHWAKARMA D K, KAPOOR R, DHIMAN A. Unified framework for human activity recognition: an approach using spatial edge distribution and transform[J]. AEU-International Journal of Electronics and Communications, 2016, 70(3): 341-353.
- [17] WANG Y, TRAN V, HOAI M. Evolution-preserving dense trajectory descriptors[J]. arXiv:1702.04037, 2017.
- [18] LI Y W, LI W X, MAHADEVAN V, et al. VLAD3: encoding dynamics of deep features for action recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 1951-1960.
- [19] ZHU J, ZOU W, ZHU Z. End-to-end video-level representation learning for action recognition[C]//24th International Conference on Pattern Recognition (ICPR), Aug 20-24, 2018, Beijing, China. Piscataway: IEEE Press, 2018: 645-650.
- [20] SUN Q, LIU H, MA L, et al. A novel hierarchical bag-of-words model for compact action representation[J]. Neurocomputing, 2016(174): 722-732.
- [21] IJINA E P, MOHAN C K. Human action recognition using genetic algorithms and convolutional neural networks[J]. Pattern Recognition, 2016(59): 199-212.
- [22] MAHASSENI B, TODOROVIC S. Regularizing long short term memory with 3D human-skeleton sequences for action recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 3054-3062.
- [23] ALHARBI N, GOTOH Y. A unified spatio-temporal human body region tracking approach to action recognition[J]. Neurocomputing, 2015(161): 56-64.
- [24] MAHASSENI B, TODOROVIC S. Regularizing long short term memory with 3D human-skeleton sequences for action recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 3054-3062.
- [25] ZHANG X, BAO Y, ZHANG F, et al. Qiniu submission to Activity Net challenge 2018[J]. 2018 Computer Vision and Pattern Recognition Challenge, arXiv: 1806.04391, 2018.
- [26] LI Y, XU Z, WU Q, et al. Submission to moments in time challenge 2018[J]. 2018 Computer Vision and Pattern Recognition Challenge, arXiv: 1808.03766, 2018.
- [27] 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.
- [28] 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.
- [29] WANG H, SCHMID C. Action recognition with improved trajectories[C]//18th IEEE International Conference on Computer Vision (ICCV), Dec 1-8, 2013, Sydney, Australia. Piscataway: IEEE Press, 2013: 3551-3558.
- [30] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [31] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//25th Annual Conference on Neural Information Processing Systems, Dec 3-6, 2012, Lake Tahoe, USA. Massachusetts: MIT Press, 2012: 1106-1114.



- [32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations(ICLR), May 7-9, 2015, San Diego, USA. New York: ACM Press, 2015: 1-14.
- [33] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jun 26-Jul 1, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 770-778.
- [34] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//28th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jun 7-12, 2015, Boston, USA. Piscataway: IEEE Press, 2015: 7-12.
- [35] ARIF S, WANG J, HASSAN U T, et al. 3D-CNN-based fused feature maps with LSTM applied to action recognition[J]. *Future Internet*, 2019, 11(2): 42.
- [36] JI S, 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.
- [37] NG Y H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification[C]//28th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jun 7-12, 2015, Boston, USA. Piscataway: IEEE Press, 2015: 4694-4702.
- [38] LIU Z, HU H F. Spatiotemporal relation networks for video action recognition[J]. *IEEE Access*, 2019(7): 14969-14976.
- [39] BACCOUCHE M, MAMALET F, WOLF C, et al. Sequential deep learning for human action recognition[C]//2nd International Conference on Human Behavior Understanding(HBU), Nov 16-16, 2011, Amsterdam, Netherlands. Berlin: Springer Verlag, 2011: 29-39.
- [40] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long-term recurrent convolutional networks for visual recognition and description[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 39(4): 677-691.
- [41] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: evolution of optical flow estimation with deep networks[C]//30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jul 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 1467-1655.
- [42] FISCHER P, DOSOVITSKIY A, ILG E, et al. FlowNet: learning optical flow with convolutional networks[C]//20th IEEE International Conference on Computer Vision(ICCV), Dec 11-18, 2015, Santiago, Chile. Piscataway: IEEE Press, 2015: 2758-2766.
- [43] YE H, WU Z, ZHAO R W, et al. Evaluating Two-Stream CNN for Video Classification[C]//5th ACM on International Conference on Multimedia Retrieval(ICMR), Jun 23-26, 2015, Shanghai, China. New York: ACM, 2015: 435-442.
- [44] WU Z, WANG X, JIANG Y G, et al. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification[C]//23rd ACM Multimedia Conference, Oct 26-30, 2015, Brisbane, Australia. New York: ACM Press, 2015: 461-470.
- [45] WU Z, JIANG Y G, WANG X, et al. Multi-stream multi-class fusion of deep networks for video classification[C]//24th ACM Multimedia Conference, Oct 15-19, 2016, Amsterdam, UK. New York: ACM Press, 2016: 791-800.
- [46] LONG X, GAN C, MELO G D, et al. Attention clusters: Purely attention based local feature integration for video classification[C]//31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 18-22, 2018, Salt Lake, USA. Piscataway: IEEE Press, 2018: 7834-7843.
- [47] JIANG Y G, WU Z, TANG J, et al. Modeling multimodal clues in a hybrid deep learning framework for video classification[J]. *IEEE Transactions on Multimedia*, 2018, 20(11): 3137-3147.
- [48] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//28th Annual Conference on Neural Information Processing Systems(NIPS), Dec 8-13, 2014, Montreal, Canada. Massachusetts: MIT Press, 2014: 568-576.
- [49] FEICHTENHOFER, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jun 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 1933-1941.
- [50] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//14th European Conference on Computer Vision(ECCV), Oct 8-16, 2016, Amsterdam, Netherlands. Berlin: Springer Verlag, 2016: 20-36.
- [51] LAN Z, ZHU Y, HAUPTMANN A G. Deep local video feature for action recognition[C]//30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jul 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 1219-1225.
- [52] ZHOU B, ANDONIAN A, TORRALBA A. Temporal relational reasoning in videos[C]//15th European Conference on Computer Vision(ECCV), Sep 8-14, 2018, Munich, Germany. Berlin: Springer Verlag, 2018: 831-846.
- [53] DIBA A, SHARMA V, VAN GOOL L. Deep temporal linear encoding networks[C]//30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), July 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 1541-1550.
- [54] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatio-temporal features with 3D convolutional networks[C]//20th IEEE International Conference on Computer Vision(ICCV), Dec 11-18, 2015, Santiago, Chile. Piscataway: IEEE Press, 2015: 4489-4497.
- [55] SUN L, JIA K, YEUNG D Y, et al. Human action recognition using factorized spatio-temporal convolutional networks[C]//

- 20th IEEE International Conference on Computer Vision (ICCV), Dec 11-18, 2015, Santiago, Chile. Piscataway: IEEE Press, 2015: 4597-4605.
- [56] QIU Z, YAO T, MEI T. Learning Spatio-temporal representation with pseudo-3D residual networks[C]//22nd IEEE International Conference on Computer Vision (ICCV), Oct 22-29, 2017, Venice, Italy. Piscataway: IEEE Press, 2017: 5534-5542.
- [57] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3D ConvNets: new architecture and transfer learning for video classification[J]. arXiv: 1711.08200, 2017.
- [58] CARREIRA J, ZISSERMAN A. Quo Vadis, action recognition? A new model and the kinetics dataset[C]//30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 6299-6308.
- [59] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 18-22, 2018, Salt Lake, USA. Piscataway: IEEE Press, 2018: 6450-6459.
- [60] FAN L, HUANG W, GAN C, et al. End-to-end learning of motion representation for video understanding[C]//31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 18-22, 2018, Salt Lake, USA. Piscataway: IEEE Press, 2018: 6016-6025.
- [61] ZHU W, HU J, SUN G, et al. A key volume mining deep framework for action recognition[C]//29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 1991-1999.
- [62] KAR A, RAI N, SIKKA K, et al. AdaScan: adaptive scan pooling in deep convolutional neural networks for human action recognition in Videos[C]//30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 5699-5708.
- [63] ZHU Y, LAN Z, NEWSAM S, et al. Hidden two-stream convolutional networks for action recognition[C]//14th Asian Conference on Computer Vision (ACCV), Dec 2-6, 2018, Perth, Australia. Berlin: Springer Verlag, 2018: 363-378.
- [64] WANG L, XIONG Y, WANG Z, et al. Towards good practices for very deep two-stream ConvNets[J]. Computer Science, arXiv: 1507.02159, 2015.
- [65] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal residual networks for video action recognition[C]//30th Conference and Workshop on Neural Information Processing Systems (NIPS), Dec 5-10, 2016, Barcelona, Spain. [S.l.:s.n.], 2016: 3476-3484.
- [66] WANG Y, LONG M, WANG J, et al. Spatiotemporal pyramid network for video action recognition[C]//30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 2097-2106.
- [67] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal multiplier networks for video action recognition [C]//30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 7445-7454.
- [68] OUYANG X, XU S J, ZHANG C Y, et al. A 3D-CNN and LSTM based multi-task learning architecture for action recognition[J]. IEEE Access, 2019(7): 40757-40770.
- [69] WANG L, QIAO Y, TANG X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]//28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 7-12, 2015, Boston, USA. Piscataway: IEEE Press, 2015: 4305-4314.
- [70] VAROL G, LAPTEV I, SCHMID C. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 1510-1517.
- [71] LEV G, SADEH G, KLEIN B, et al. RNN fisher vectors for action recognition and image annotation[C]//14th European Conference on Computer Vision (ECCV), Oct 8-16, 2016, Amsterdam, Netherlands. Berlin: Springer Verlag, 2016: 833-850.
- [72] BILEN H, FERNANDO B, GAVVES E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2799-2813.

[作者简介]



赵朵朵 (1995—)，女，杭州电子科技大学通信工程学院硕士生，主要研究方向为图像处理与人工智能等。

章坚武 (1961—)，男，博士，杭州电子科技大学通信工程学院教授、博士生导师，中国电子学会、中国通信学会高级会员，浙江省通信学会常务理事，主要研究方向为移动通信、多媒体信号处理与人工智能、通信网络与信息安全。

郭春生 (1971—)，男，博士，杭州电子科技大学通信工程学院副教授、硕士生导师，主要研究方向为视频分析与模式识别。

周迪 (1975—)，男，浙江宇视科技有限公司高级工程师、宇视研究院院长，主要研究方向为视频安全、人工智能等。

穆罕默德·阿卜杜·沙拉夫·哈基米 (1991—)，男，杭州电子科技大学博士生，主要研究方向为图像处理与人工智能。