基于深度学习的视频中人体动作识别进展综述

罗会兰[1],童 康[1],孔繁胜[2]
(1. 江西理工大学信息工程学院,江西赣州 341000; 2. 浙江大学计算机科学技术学院,浙江杭州 310027)

摘 要: 视频中的人体动作识别是计算机视觉领域内一个充满挑战的课题. 不论是在视频信息检索、日常生活安全、公共视频监控,还是人机交互、科学认知等领域都有广泛的应用. 本文首先简单介绍了动作识别的研究背景、意义及其难点,接着从模型输入信号的类型和数量、是否结合了传统特征提取方法、模型预训练三个维度详细综述了基于深度学习的动作识别方法,及比较分析了它们在UCF101和HMDB51这两个数据集上的识别效果. 最后分别从视频预处理、视频中人体运动信息表征、模型学习训练这三个角度对未来动作识别可能的发展方向进行了论述.

关键词: 动作识别; 综述; 卷积神经网络; 深度学习
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2019)05-1162-12
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.05.025

The Progress of Human Action Recognition in Videos Based on Deep Learning: A Review

LUO Hui-lan[1], TONG Kang[1], KONG Fan-sheng[2]
(1. *School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi* 341000, *China*;
2. *School of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*)

Abstract: Human action recognition in videos is a challenging topic in the field of computer vision. It is widely not only used in video information retrieval, daily life security, public video surveillance, but also human-computer interaction, scientific cognition and other fields. First, the research background, research significance and difficulties of action recognition are briefly introduced, and then the deep learning model based action recognition methods are comprehensively reviewed from three different aspects: the types and numbers of input signals, the combination with traditional feature extraction methods, and the pre-trained datasets. Furthermore, the performances of some typical methods on UCF101 and HMDB51 datasets are overviewed and analyzed. Last the possible future research directions are discussed from three perspectives: the video data preprocessing, the video human motion feature representation, and the model training.

Key words: action recognition; review; convolutional neural network; deep learning

1 引言

随着网络多媒体的快速发展以及视频获取设备的日渐普及,越来越多的视频被共享. 如何理解和分析这些海量涌现的视频数据具有重大的理论及应用价值. 视频中的人体动作识别已经成为计算机视觉领域研究的热点,它通过对图像序列处理分析,学习并理解其中人的动作和行为,剖析人体运动模式,建立视频内容和动作类型之间的映射关系,使得计算机能够像人类一样去"看"视频,从而"理解"视频. 因此,不论是从视频信息检索、日常生活安全、公共视频监控,还是人机交互、科学认知等角度,对视频中人体动作识别进行研究都具有重要的学术价值和应用价值.

视频中的人体动作识别主要分为两个步骤,先是视频中人体动作的特征表示,然后是对这些特征进行理解并最终分类. 视频图像的特征表达是视频动作识

收稿日期:2018-07-09;修回日期:2018-10-09;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 61462035, No. 61862031);江西省自然科学基金(No. 20171BAB202014);江西省青年科学家培养对象计划资助(No. 20153BCB23010)

别的重点,可以分成基于模型的表示和不基于模型的表示,其中后者又可分为局部特征表示和全局特征表示. 近年来,卷积神经网络、递归神经网络等深度学习模型理论为特征表示提供了新的方法. 传统的动作识别方法是将特征的提取与后续动作识别的训练分成二个独立的过程,在获得动作视频的特征表示后输入机器学习算法进行训练,实现最终的分类与识别. 而基于深度学习模型的端到端方法是将特征提取表达与后续的全连接网络层进行统一训练和学习,实现特征提取和分类的无缝连接. 目前视频中的人体动作识别研究存在的挑战和难点主要有以下几点[1].

(1)视频数据源的不稳定

视频采集摄像头的抖动、环境光照和温度的改变、动态变化的嘈杂背景这些因素都会对特征提取算法的选择以及算法的计算结果产生较大的影响.

(2)动作的类内差异性和类间相似性

不同人做同一类动作,即使同一个人做同一类动作,由于个体差异、动作快慢、环境及背景等不同,在视频中的表现可能会有很大不同. 而不同类的动作又可能表现出很相似的特征. 随着动作类别数量的增加,不同动作姿态之间重叠程度随着表示空间的细分而不断增加,即同类动作之间经常具有大的类内散度,而不同动作类之间具有小的类间散度,这对动作的识别也是一个挑战.

(3)连续动作的分割和长时视频中动作的识别

人体行为往往由一连串的动作构成,而动作之间却没有明显的边界指示. 现有的动作识别方法大多数是对已经从时间域分割好的视频片段来进行分类,对于长时视频中发生的多个动作并不能够很好的识别,以及对于事件发生的开始帧和结束帧也不能很好的定位.

(4)大量训练数据的标注

人体动作识别的研究经历了从受限场景中简单动作的识别到电影中的动作识别,再到自然生活场景中的人体动作识别. 如何对收集到的视频数据进行标注是一个问题. 人工标注的方法费时又耗力,显然不可取,迫切需要对视频数据自动标注的工具或技术.

视频中的人体动作识别研究一直广受研究者的青睐,近年来取得了许多进展. 显然,早期的动作识别研究综述工作[2~4]并没有包含对最近基于复杂深度模型动作识别方法的论述. 胡琼等人[1]从特征提取的方法、动作识别的方法、相关国际竞赛和常用数据库等方面详细阐述了基于视觉的人体动作识别领域的研究现状、研究难点以及可能的发展方向. 文献[5,6]对人体动作识别的相关数据集进行了总结和归纳. 根据数据集的数据特点和获取方式的不同,朱红蕾等人在文献[7]中对人体行为识别公开数据集进行分类归纳,旨在引导研究者选取合适的基准数据集来验证其算法的性能. 文献[8]明确讨论了现有手动特征提取的动作识别方法以及基于学习表示的动作识别方法的优点和局限性. Samitha Herath 等人[9]从手工表示的动作识别开创性方法开始讨论,然后转向基于深度学习领域方法的论述. 类似地,文献[10]从基于人工提取特征的动作识别方法到基于深度学习表示的动作识别方法进行了非常全面的综述,而且对可用的数据集以及动作识别的应用和未来的研究方向进行了重要讨论. 文献[11]在单视角、多视角、RGB 深度视频这三种类型数据集上对基于深度学习技术的动作识别算法进行了论述. Guangle Yao 等人[12]从如何利用卷积网络学习时间信息的角度,综述了基于卷积神经网络的动作识别方法. 不同于以上这些文献综述,本文则是从深度模型的输入信号类型及数量、深度模型特征和传统手动特征结合方法、预训练方法这几个角度来详细综述基于深度模型的视频人体动作识别研究方法,并在 UCF101 和 HMDB51 这两个数据集上分析比较了一些有代表性的识别方法,并通过三个不同角度:视频预处理、视频中人体运动信息表征、模型学习训练对未来动作识别可能的发展方向进行了阐述. 旨在对当前基于深度模型的视频动作识别方法进行对比分析,得到一些研究方向启发.

2　基于不同输入类型的深度模型动作识别算法综述

2.1　输入信号的类型

大部分的文献都会直接使用视频帧的 RGB 信号作为深度模型的输入[13~15]. Du Tran 等人[13]分别使用三维卷积和三维池化直接对 RGB 信号流进行卷积和池化. 同样地,Christoph Feichtenhofer 等人[15]也用到了三维卷积和三维池化对输入的 RGB 信号流进行处理. 其次,在视频动作识别中,广泛使用的输入信号类型是光流(Optical Flow),它指的是时变图像中模式的运动速度. 当物体在运动时,图像上对应点的亮度模式也在运动,这种亮度模式的运动表观就是光流. Karen Simonyan 和 Andrew Zisserman 在文献[16]中首次将多帧堆叠的光流信号运用于二流网络中,开创了时间流卷积网络的先河. 受改进稠密轨迹[17]的启发,Limin Wang 等人[18]通过估计单应性矩阵,然后补偿摄像机的运动来提取扭曲的光流(Warped Optical Flow),并将其作为时间流卷积网络的信号输入. 为了使卷积神经网络能更好的提取帧之间的时间信息,Shuyang Sun 等人[19]提出了光流引导特征(Optical Flow guided Feature,OFF). 通过直接计算深度特征图的像素级别的时空梯度,光流引导特征可以嵌入到现有的任何卷积神经网络中,OFF(RGB)表示使用光流引导特征对输入 RGB 信号的处理,OFF(Optical Flow)表示使用光流引导特征对输入光

流信号的处理. Bowen Zhang 和 Limin Wang 等人为了缓解光流计算的高成本,在文献[20]中提出了运动矢量(Motion Vector),它可以用比光流更低的计算成本直接从视频的解码过程中获得.

由 RGB 输入信号和光流输入信号构成的卷积神经网络能够分别学习到视频的空间信息和局部时间信息,为了能够学习视频的全局时间信息,Liangliang Wang 等人[21]提出了运动堆叠的差分图像(Motion Stacked Differece Image,MSDI),并将其作为全局时间流的输入信号. 运动堆叠的差分图像是通过在整个动作序列中,以相邻图像帧之间的绝对差分形式融合每个局部动作特征来建立的. 大多数现存的特征或描述符不能够有效捕获运动信息,尤其是长时间的运动信息. 为此,Yemin Shi 等人[22]提出了一个名为连续深度轨迹描述符(sequential Deep Trajectory Descriptor,sDTD)的长期运动描述符,该描述符的提取过程由三个部分组成:(i)提取简化的稠密轨迹(ii)将从每个视频中提取的这些轨迹转换成一系列的轨迹纹理图像(iii)使用卷积神经网络和递归神经网络来学习连续深度轨迹描述符. 为了将视频中的动态信息表示成一张图像,Hakan Bilen 等人[23]提出了动态图像(Dynamic Image)的概念,通过对视频帧应用排序池化[24]将视频表达成单帧图像,从而可以利用图像处理模型进行特征提取和学习. 随后,他们将此方法在光流序列中扩展,在文献[25]中提出了动态光流(Dynamic Optical Flow)的概念,通过应用近似排序池化将光流序列表示成单帧光流特征图,并将其应用于动作识别中.

2.2 输入信号流的个数

根据输入信号流的数量,当前基于深度模型的动作识别方法可以分为单流、双流及多流网络模型方法.

2.2.1 单流动作识别模型

在单流动作识别模型中,使用最广泛的是三维卷积神经网络. 三维卷积的概念最初是 Shuiwang Ji 等人在文献[26]中提出的,三维卷积是通过将多个连续帧堆叠形成的立方体和一个三维核进行卷积获得. 基于三维卷积,Shuiwang Ji 等人提出了一个包含 1 个硬连接层、3 个卷积层、2 个子采样层和 1 个全连接层的三维卷积网络结构,以在时间和空间维度提取视频特征. 在文献[26]的基础上,Du Tran 等人[13]提出了深度三维卷积神经网络,该方法直接利用深度三维卷积网络中的三维卷积和三维池化对 RGB 视频进行处理,并利用大规模有监督视频数据集进行训练获得 C3D(Convolutional 3D)模型,网络结构如图 1 所示.

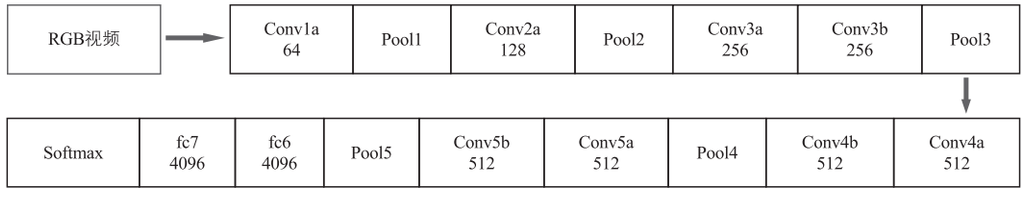

图1 三维卷积网络结构图

该三维卷积网络有 8 个三维卷积层、5 个最大三维池化层、2 个全连接层和一个 Softmax 输出层. 图 1 中三维卷积层方框里的数字表示滤波器的数量,采用的三维卷积核大小为 3×3×3. 第一个三维池化层池化核大小为 1×2×2,其他的池化核大小为 2×2×2,全连接层的维度为 4096 维. 随后,Du Tran 等人[27]将残差网络和三维卷积相结合,提出了一个深度三维残差卷积网络(3-Dimensional Residual ConvNet). 和三维卷积网络相似,他们利用三维残差卷积网络对 RGB 视频进行处理,并在数据集 Sports-1M[14]上训练获得 Res3D 模型,它比 C3D 模型小 2 倍且运行速度更快. 时空信息融合具有高的训练复杂度以及三维卷积具有高的计算和内存成本阻碍了当前三维卷积神经网络的发展,Yizhou Zhou 等人[28]提出了一个混合的卷积管(Mixed Convolutional Tube,MiCT),它将二维卷积神经网络和三维卷积模块集成在一起,产生更深入、更丰富的特征图,同时减小了每一轮时空融合的训练复杂度. 基于 MiCT 堆叠形成的深度三维网络 MiCTNet 能够很好的利用 RGB 视频中人类动作的时空信息,并且获得了比原三维卷积神经网络更优的性能. 在文献[29]中,Unaiza Ahsan 等人提出了使用生成对抗网络(Generative Adversarial Networks,GAN)对 RGB 视频进行训练,并将此过程作为识别的初始化步骤,以此来学习视频中的动作表示,同时获得了不错的识别效果. 此外,也有许多研究者探索深度卷积神经网络的视频输入新方式,直接利用图像处理深度模型到视频任务中.

2.2.2 双流动作识别模型

二流架构是由神经科学[30]的二流假说所激发的. 假设视觉皮层有两种途径:一种途径是腹侧通路,它对目标的形状和颜色响应空间特征;另一种途径是背侧通路,它对目标转换和由运动引起的空间关系很敏感. 受此启发,Karen Simonyan 和 Andrew Zisserman 在文献[16]中提出了二流卷积神经网络并将其用于动作识别,结构图如图 2 所示.

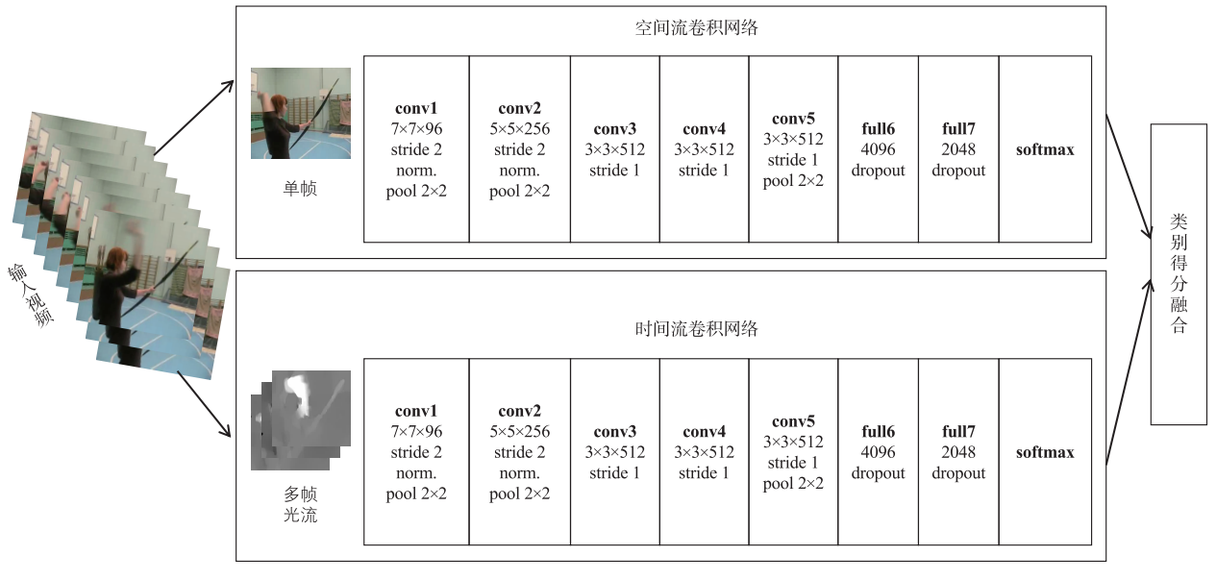

图2　二流网络结构图

该架构包含空间流部分和时间流部分. 空间流部分以单个图像帧为输入,然后经过带有 5 个卷积层和 2 个全连接层的卷积神经网络,形成了空间流卷积网络;时间流部分以 10 帧光流的堆叠作为输入,同样经过带有 5 个卷积层和 2 个全连接层的卷积神经网络,形成了时间流卷积网络. 空间流网络通过随机采样视频帧进行训练,时间流网络通过随机采样视频片段进行训练,两个网络流独立进行训练. 测试阶段,对视频进行等间隔采样,通过取均值分别汇总时间流网络和空间流网络的 softmax 得分,然后进行双流的融合. 作者提出了平均融合方法和把 softmax 得分作为特征训练线性支持向量机的融合方法,这种融合方法属于分类器级的融合. 虽然分类器级的融合实现相对简单,但是并不能够很好的融合时间流和空间流信息. 为此,Christoph Feichtenhofer 等人[15]提出了一种新的双流融合方法,通过在卷积层之后进行融合,不仅实现了从分类器级到特征级融合的转变,而且获得了更优的性能. 二流卷积神经网络中使用的光流输入计算十分耗时,严重阻碍了二流结构的处理速度,为此 Bowen Zhang 等人[20]提出了一种高性能实时动作识别框架,从输入的压缩视频中提取 RGB 图像和运动矢量. 运动矢量缺乏精细准确的运动信息,直接用它来替代光流会降低时间卷积神经网络的识别性能,并且运动矢量表示图像的运动模式和光流也很相似,故此增强的运动矢量被提出,通过三种知识迁移的技巧,利用光流卷积神经网络学到的知识来强化运动矢量为输入的时间卷积神经网络,并与 RGB 图像为输入的空间卷积神经网络相结合构成新的二流结构,不仅获得了和原始二流网络相媲美的效果,而且处理速度比原始二流网络快了 27 倍.

视频动作识别的训练数据集相较于图像领域的 ImageNet[31]数据集来说相对较小,在训练时更容易出现过拟合. 因此相比于图像识别领域的深度模型(如 VGGNet[32]和 GoogLeNet[33]),视频中的动作识别采用的二流卷积网络结构相对较浅,表达能力受到网络深度的限制. 为了解决这个问题,Limin Wang 等人[34]提出了极深的二流卷积网络,通过利用 ImageNet 上预训练好的 VGG-16 和 GoogLeNet 网络对单帧图像和 10 帧光流堆叠的输入进行处理,并设计了 4 个好的训练技巧:一是对时间和空间网络进行预训练,二是使用小的学习率,三是使用更多的数据增强技术,四是利用高的 dropout 率. Christoph Feichtenhofer 等人[35]在二流方法的基础上,将图像分类数据集上预训练好的残差网络[36]应用到动作识别中,并且将原二流卷积网络中的空间流和时间流改名为外观流和运动流. 外观流和运动流在训练时使用了随机水平翻转,但没有使用 RGB 颜色抖动. 此外通过在两个流之间注入残差连接来对模型进行联合微调. 为了减小最终时空残差网络模型的过拟合,在一个视频中,从 5 到 15 帧的时间步长中随机采样 5 个输入来训练时空残差网络,最后对时间流和空间流全连接层的预测求平均来识别分类. 随后,Christoph Feichtenhofer 等人[37]探索了许多连接外观流和运动流的方法,又提出了乘法交互的跨流残差连接,通过将恒等映射核作为时间滤波器注入到网络模型中,以捕获长期的时间依赖性,这种新的时空乘法网络结构在动作识别上获得了好的性能.

除了使用卷积神经网络之外,长短时记忆网络也被应用于视频动作识别中. Joe Yue-Hei Ng 等人[38]为了处理完整长度的视频,分别将 120 个视频帧和 30 个光

流帧输入卷积神经网络,然后分别用特征池化和长短时记忆单元对卷积神经网络输出的特征进行特征聚合,最后融合分类得分进行动作识别. Lin Sun 等人[39]通过学习记忆细胞的独立隐藏状态转换来扩展长短时记忆单元,提出了 Lattice-LSTM 的网络机制,并将该机制运用于二流卷积神经网络结构中,通过多模式训练的方式对网络进行训练. Yilin Wang 等人[40]提出了一个混合深度框架,将分层结构与联合注意力模型合并到二流卷积网络中,每一个流都使用来自长短时记忆单元编码的特征计算注意力的权重,并将输入特征的加权平均添加进长短时记忆单元,如此经过多个长短时记忆单元的处理来对视频中的动作进行识别. Yunbo Wang 等人[41]将紧凑双线性算子和时空注意力的方法引入到视频中的人体动作识别任务中,提出了一种新的端到端的时空金字塔结构,使得由 RGB 图像主导的空间线索和以光流为主导的时间线索相互促进,以此提高识别的性能. 为了尽可能学习到视频中包含的完整运动信息,Gul Varol 等人[42]使用具有长时间卷积(Long-term Temporal Convolution,LTC)的神经网络来学习视频表示,一定程度上扩宽了时间范围,从而提升动作识别的准确率. Limin Wang 等人[18]为了利用整个视频的视觉信息来进行视频级别的预测,提出了时间分割网络概念,从整个视频中稀疏采样片段,每个片段各自产生动作类的初步预测,然后将这些片段之间的共识作为最终视频级别的预测. Ali Diba 等人[43]提出了时间线性编码(Temporal Linear Encoding,TLE)概念,它可以将整个视频编码成一个紧凑的特征表示,以此来学习语义和一个特征空间,通过将时间线性编码分别嵌入到 RGB 帧为输入的卷积网络和 10 帧光流堆叠为输入的卷积网络中形成新的层,最后对其预测得分求平均进行动作识别.

2.2.3 多流动作识别模型

为了尽可能多地获取到动作视频中的特征信息,Liangliang Wang 等人[21]提出了一个全局时空三流卷积神经网络结构. 如图 3 所示,他们利用单帧图像、10 帧光流堆叠以及运动堆叠的差分图像作为深度卷积神经网络的输入,训练获得空间流、局部时间流和全局时间流特征. 对三流卷积神经网络融合时,他们先对学到的特征进行 PCA-Whitening 操作,然后对这些特征数据进行 soft-VLAD(soft Vector of Locally Aggregated Descriptor)矢量编码,最后使用支持向量机进行分类. 学习运动信息的时空表征是动作识别的关键,Yemin Shi 等人[22]提出了能够有效捕捉长期运动信息的连续深度轨迹描述符(sequential Deep Trajectory Descriptor,sDTD),分别读取 16 帧连续深度轨迹描述符、16 个图像帧以及 16 个光流帧,并利用卷积神经网络和递归神经网络进行训练,分别形成连续深度轨迹描述符流、空间流和时间流,最后将它们全连接层的预测输出进行后期融合以进行分类识别. Hakan Bilen 等人[25]利用动态图像的概念,分别使用排序池化和近似排序池化对 RGB 图像和光流进行编码,并将其输入卷积神经网络训练得到 RGB 动态图像流网络和动态光流网络,结合原始 RGB 流网络和光流网络形成四流网络结构. 该四流网络都是单独进行训练. 测试时,用四流网络输出得分的均值来预测动作类. Shuyang Sun 等人[19]提出了一种新的紧凑运动表示,这种表征被称为光流引导特征(Optical Flow guided Feature,OFF),它能够快速从卷积神经网络中提取鲁棒的时空信息. 通过求和汇总 RGB、OFF(RGB)、Optical Flow、OFF(Optical Flow)为输入的四流卷积网络的得分进行动作识别.

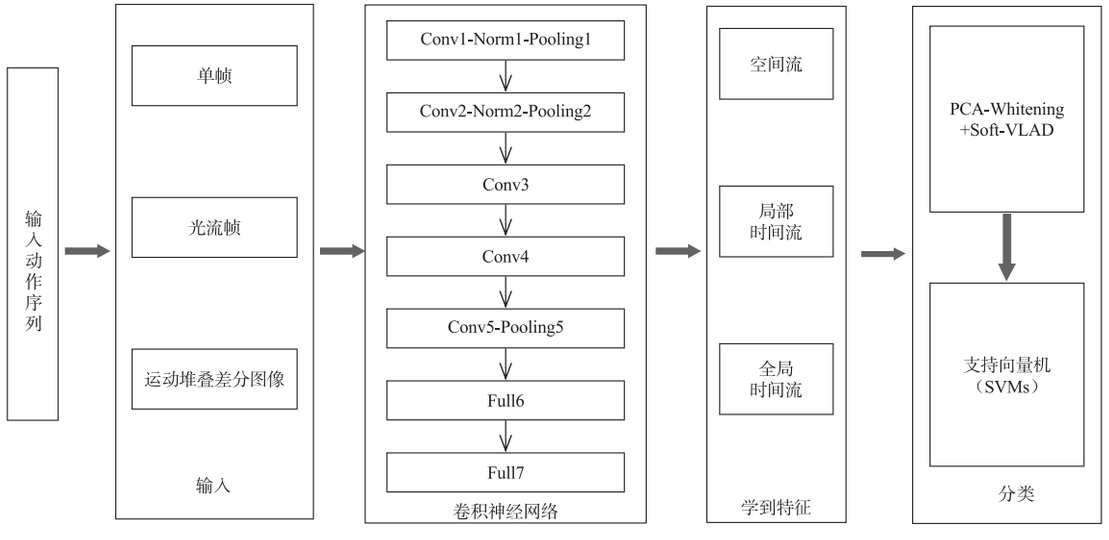

图3 三流网络结构图

2.3 不同输入类型的深度模型识别算法性能比较分析

当前最流行的用于动作识别算法性能比较的数据集是 UCF101[44] 数据集和 HMDB51[45] 数据集. UCF101[44] 数据集包含 13320 个来自 101 个动作类别的视频片断. 该数据集的 101 个类别可以分为五大类:体育运动、乐器演奏、人与人之间的交互、身体运动、人与对象的交互. 由于该数据集来源于现实环境,包含杂乱背景、相机抖动、遮挡、不同光照条件等各种因素的影响,故此该数据集有一定的挑战性,同时引起了众多学者的关注. HMDB51[45] 数据集是一个大而真实的视频集合,它包含 51 个动作类别,每个类别至少含有 101 个视频片段,总共涵盖了 6766 个视频片断. 这些视频片断主要来源于电影,只有一小部分来自公共数据库,并且每一个片断都包含一个人类活动. 该数据集的行为类别包括普通面部动作、操纵对象面部动作、一般身体运动、与对象交互运动、与人交互运动等五种类型. HMDB51 数据集来源不同,并伴有遮挡、相机移动、复杂背景、光照条件变化等诸多因素的影响,故相较于 UCF101 数据集,HMDB51 数据集极具挑战性,识别难度更大.

表 1 给出了不同输入类型下,不同算法在 UCF101 和 HMDB51 数据集上的识别准确率对比. 报告的识别准确率直接来源于相应原文献. 为了表示方便,在表 1 "输入"栏中,A 表示基于"A"的单流输入、A + B 表示基于"A + B"的双流输入、A + B + C 表示基于"A + B + C"的三流输入、A + B + C + D 表示基于"A + B + C + D"的四流输入. 此外,Motion Vector、Dynamic Image、Dynamic Optical Flow、Warped Optical Flow、Optical Flow 在表中分别简化表示为 MV、DI、DOF、WOF 和 OF.

从表 1 中可以看出,相较于其他输入,RGB、Optical Flow 以及 RGB + Optical Flow 是输入中被广泛使用的类型. 在 RGB 单独信号流输入的单流动作识别模型中,Hakan Bilen 等人[25] 使用 ResNeXt-50 模型的 RGB 网络流取得了较好的结果,在 UCF101 和 HMDB51 数据集上分别达到了 87.6% 和 53.5% 的准确率. 在 Optical Flow 单独信号流输入的单流动作识别模型中,Christoph Feichtenhofer 等人[37] 使用 ResNet-50 模型的运动流在 UCF101 数据集上展现了好的结果,准确率达到 87.0%. Gul Varol 等人[42] 使用具有长时间卷积的神经网络来学习视频表示,通过 LTC-CNN 模型来扩宽时间范围,在 HMDB51 数据集上获得了 59.0% 的准确率. 此外,基于 Dynamic Image 的单独信号流输入的单流动作识别模型和基于 Dynamic Optical Flow 的单独信号流输入的单流动作识别模型在 UCF101 和 HMDB51 数据集上的性能要优于连续深度轨迹描述符(sDTD)为输入的单流动作识别模型. 在基于双输入信号流的双流动作识别模型中,基于 RGB + Optical Flow 信号流输入的双流动作识别模型在 UCF101 和 HMDB51 数据集上的

表 1　不同输入类型下动作识别的性能对比

输入	方法	UCF101	HMDB51
RGB	Slow fusion[14]	65.4%	–
	Spatial stream[16]	73.0%	40.5%
	Spatial HAN[40]	75.1%	47.7%
	Appearance Stream(ResNet-50)[37]	82.3%	48.9%
	C3D(1 net)[13]	82.3%	–
	LTC[42]	82.4%	–
	Spatial stream[22]	82.9%	–
	RGB stream(ResNeXt-50)[25]	87.6%	53.5%
sDTD	sDTD[22]	71.7%	41.1%
OF	Temporal stream[22]	75.3%	–
	Temporal stream[16]	83.7%	54.6%
	OF stream(ResNeXt-50)[25]	84.9%	55.8%
	LTC[42]	85.2%	59.0%
	Temporal HAN[40]	85.4%	58.3%
	Motion stream(ResNet-50)[37]	87.0%	55.8%
DI	DI stream(ResNeXt-50)[25]	86.6%	57.3%
DOF	DOF stream(ResNeXt-50)[25]	86.6%	58.9%
OF + sDTD	Temporalstream + sDTD[22]	82.5%	–
RGB + MV	RGB-CNN + EMV-CNN[20]	86.4%	–
RGB + OF	Two-stream(SVM fusion)[16]	88.0%	59.4%
	LSTM[38]	88.6%	–
	ST-stream[22]	90.0%	58.4%
	Very deep two-stream[34]	91.4%	–
	LTC[42]	91.7%	64.8%
	Two-stream(conv fusion)[15]	92.5%	65.4%
	HAN[40]	92.7%	64.3%
	STRN[35]	93.4%	66.4%
	Lattice-LSTM[39]	93.6%	66.2%
	Two-stream(SI + OF)[25]	93.9%	67.5%
	TSN(2 modalities)[18]	94.0%	68.5%
	STMN[37]	94.2%	68.9%
	STPN(BN-Inception)[41]	94.6%	68.9%
	Two-stream(RGBTLE + Flow TLE)[43]	95.6%	71.1%
OF + DOF	Two-stream(OF + DOF)[25]	89.1%	62.6%
RGB + sDTD	Spatialstream + sDTD[22]	89.7%	–
RGB + DI	Two-stream(SI + DI)[25]	90.6%	61.3%
RGB + OF + MSDI	Three-stream(SVM fusion)[21]	89.7%	61.3%
RGB + OF + sDTD	ST-stream + sDTD[22]	92.1%	63.7%
RGB + OF + WOF	TSN(3 modalities)[18]	94.2%	69.4%
RGB + OF + DI + DOF	Four-stream(SI + DI + OF + DOF)[25]	95.0%	71.5%
RGB + OFF(RGB) + OF + OFF(OF)	Four-stream[19]	96.0%	74.2%

性能要优于基于 RGB + sDTD、RGB + Motion Vector、RGB + Dynamic Image、Optical Flow + sDTD、Optical Flow + Dynamic Optical Flow 信号流输入的双流动作识别模型. 在 RGB + Optical Flow 信号流输入的双流动作识别模型中,Ali Diba 等人[43]采用基于双输入 RGB TLE + Flow TLE 的模型在数据集 UCF101 和 HMDB51 上的识别率分别达到了 95.6% 和 71.1%,由此可见通过时间线性编码(Temporal Linear Encoding,TLE)对 RGB 和 Optical Flow 进行处理后能够更好的提取时空特征,从而获得更好的识别效果. 在基于多输入信号流的多流动作识别模型中,Shuyang Sun 等人[19]采用的基于多输入 RGB + OFF(RGB) + Optical Flow + OFF(Optical Flow)的网络模型在数据集 UCF101 和 HMDB51 上获得了最好的结果,识别率分别达到了 96.0% 和 74.2%. 从表 1 还可以看出,双输入、多输入构成的双流和多流动作识别模型的识别效果要明显优于单输入构成的单流动作识别模型. 双流动作识别模型在 UCF101 和 HMDB51 数据集上平均准确率相比较于单流动作识别模型分别提升了 10.2% 和 12.6%. 而多流动作识别模型在 UCF101 和 HMDB51 数据集上平均准确率相比较于单流动作识别模型分别提升了 12.4% 和 15.4%.

3 结合深度模型和传统手动提取特征的动作识别方法综述

3.1 手动特征提取方法

传统手动特征提取的动作识别方法基本上都是先从原始输入图像帧中检测运动信息并提取底层特征,然后对动作进行建模,建立底层特征和动作行为类别的对应关系,即利用这些特征学习出一个分类器. 传统的手动特征提取方法主要有整体特征提取和局部特征提取这两大类.

整体特征表示的动作识别方法是基于对人体结构、几何形状、以及运动的全局表征的识别方法. 根据人体的几何形状进行动作识别是最直接的方法. 但该方法[46]受限于人体几何形状的建模以及运动中人体形状的柔性,因此简单的数学模型并不能够很好的描述运动过程中的人体形状. Bobick 等人的研究采用运动能量图像[47](Motion Energy Images,MEI)和运动历史图像[48](Motion History Images,MHI)来解释图像序列中人的运动[49]. 运动能量图像捕获运动发生的地方,而运动历史图像模板展示的是图像如何运动. 此外通过时空卷[48]来表示运动历史图像模板,从二维特征图扩展到三维卷,增加了视角点变化的鲁棒性. Alper Yilmaz 和 Mubarak Shah 在文献[50]中提出了根据时空卷(Space-Time Volume,STV)的不同属性来识别动作,时空卷(STV)通过在时间轴上叠加目标的轮廓来构建,一个时空卷的方向、速度和形状的改变表征了潜在的动作. 但是这两类方法都需要将运动人体从背景中分割出来,所以在复杂动态背景情况下效果不好.

局部特征表示的动作识别方法是基于局部特征的提取. 传统的基于局部特征提取的动作识别方法遵循的规则是先进行兴趣点检测,然后提取局部描述子,再对局部描述子进行聚合. 常见的兴趣点检测器有时空兴趣点(Space-Time Interest Point,STIP)[51]、Harris[52]、3D-Hessian[53]等. 局部特征描述子有方向梯度直方图[54](Histogram of Oriented Gradients,HoG)、HoG-3D[55]、光流直方图[56](Histogram of Optical Flow,HoF)、运动边界直方图[57](Motion Boundary Histogram,MBH)、局部二值模式[58](Local Binary Pattern,LBP)等. 早期的工作几乎都选择小的视频立方体,即小的视频时空卷作为兴趣点. 近年来,基于跟踪点的运动轨迹[59]进行特征的提取与描述引起了研究者的兴趣. A. Klaser 等人[60]为了更加有效的捕获运动信息,通过采样和跟踪多个尺度上每帧的稠密点来提取稠密轨迹. 他们还在每个点上提取了方向梯度直方图、光流直方图和运动边界直方图,这些特征的组合进一步提升了性能. Heng Wang 等人[17]提出了改进的稠密轨迹特征(Improving Dense Trajectories,IDT),它是对稠密轨迹的改进,通过考虑相机运动估计,然后应用特征词袋或费舍尔向量编码每个视频的最终特征表示. 改进的稠密轨迹特征在提取过程中包含几个重要的部分. 第一,这些提取的轨迹主要位于高速运动的显著区域,该区域包含了动作识别的丰富信息. 第二,在几个连续帧中相应区域的局部描述符是沿着轨迹排列的,并且受轨迹限制的采样策略能够有效处理运动速度的变化.

3.2 结合基于轨迹的特征和深度模型方法

虽然传统的动作识别方法具有高的人工计算和测试成本以及需要足够的经验和相当的运气来获取鲁棒的特征表达,但是它们对于特征提取更具有针对性. 许多研究者探索将改进的稠密轨迹特征[17]和深度模型学习方法结合,期待得到性能的提升. Christoph Feichtenhofer 等人[15]对得到的 IDT 描述符(即 HOG、HOF 和 MBH)进行费舍尔向量编码后,利用支持向量机进行分类,并将得分和深度卷积神经网络的预测得分进行均值整合,得到了性能的提升. 随后,Christoph Feichtenhofer 等人在文献[37]中再一次结合改进的稠密轨迹特征,通过将卷积网络模型的预测层输出和费舍尔矢量编码的 IDT 特征的 L2 标准化支持向量机得分求平均来进一步提升性能. Lijie Fan 等人[61]一方面将方向梯度直方图、光流直方图、运动边界直方图等局部描述子这些 IDT 特征,采用费舍尔向量进行编码并进行 L2 标准化,再通过支持向量机获得相应的得分. 另一方面,

采用卷积神经网络进行训练和测试,通过深度学习的方法获得相应模型预测输出. 最后,对这两者求平均来获得最终的结果.

深度模型算法 Dynamic-image[23]、LTC[42]、STMN[37]和 TVNet[61],它们在数据集 UCF101 上的识别准确率分别为 76.9%、91.7%、94.2% 和 94.5%,在数据集 HMDB51 上的识别准确率分别为 42.8%、64.8%、68.9% 和 71%. 而它们通过与手动 IDT 特征结合后,在数据集 UCF101 上的识别准确率分别为 89.1%、92.7%、94.9% 和 95.4%,在数据集 HMDB51 上的识别准确率分别为 65.2%、67.2%、72.2% 和 72.6%. 显然,通过与手动 IDT 特征的结合,该 4 种深度模型算法在 UCF101 和 HMDB51 这两个数据集上都获得了不同程度的性能提升,在数据集 UCF101 上识别准确率分别提升了 12.2%、1%、0.7% 和 0.9%,在 HMDB51 数据集上识别准确率分别提升了 22.4%、2.4%、3.3% 和 1.6%. 虽然在 UCF101 数据集上的性能补偿没有在 HMDB51 数据集上那么明显,但是它相对于不添加 IDT 特征的原始方法也获得了一定程度的性能提升. 在 UCF101 和 HMDB51 数据集上性能的提升,可以归因于 IDT 特征对相机运动的确切补偿,并且该补偿在 HMDB51 数据集上的表现更加明显. 数据集 HMDB51 受相机抖动、复杂背景等因素的影响要大于数据集 UCF101. 此外,数据集 HMDB51 上同一动作具有大的类内散度以及不同动作具有小的类间散度的程度要大于数据集 UCF101. 手动 IDT 特征能够更好的对这两者的影响进行补偿,这可能就是 HMDB51 数据集上平均准确率提升比 UCF101 数据集上高的原因.

4　预训练对深度模型方法的影响分析

视频中的人体动作识别训练通常分为两大类:一类是只在特定数据集上训练,另一类是利用大型公共数据集进行预训练,再在特定数据集上微调. 预训练主要有两个优点:第一,预训练可以弥补目标数据集没有足够多的训练样本,并且它是一个有效初始化卷积神经网络的方法. 第二,预训练一定程度上可以加速网络收敛的速度,减少卷积神经网络的训练时间. 常见用于预训练的数据集有:ImageNet[62]数据集、Sports-1M[14]数据集和 Kinetics[63]数据集.

ImageNet[62]数据集拥有超过 1500 万的高分辨率图像,涵盖 22000 个类别. 这些图片来源于网络并使用亚马逊土耳其机器人众包工具来标注标签. ImageNet 大规模视觉识别挑战竞赛(ILSVRC)使用了 ImageNet 的一个子集,每组大约 1000 张图像,1000 个类别. 总共大约有 120 万张训练图片,5 万张验证图片和 15 万张测试图片. Lin Sun 等人[64]提出了分解的时空卷积网络,并引入了辅助分类器层来连接空间卷积层,使用 ImageNet 数据集预训练这个辅助网络,并随机采样训练视频帧来微调. Limin Wang 等人[18]利用数据集 ImageNet 上预训练的模型来初始化以 RGB 图像为输入的空间网络,随后利用空间网络模型来初始化时间网络. Christoph Feichtenhofer 等人[35]运用 ImageNet 数据集上预训练好的 ResNet-50 模型,构建二流网络来识别动作. 此外,Christoph Feichtenhofer 等人在文献[37]中还利用了 ImageNet 上预训练好的极深 ResNet-152 模型来构建二流网络进行动作识别.

Sports-1M[14]数据集是 Google 公布的一个大型视频数据库,来自于公开的 YouTube 视频. 该数据集包含 487 种体育运动项目,共计 1133158 个视频. 该数据集中每种行为类别包含 1000 ~ 3000 个视频,其中有大约 5% 的视频带有多个标注. 该数据集包含的体育运动项目可以分为六大类:水上运动、团队运动、冬季运动、球类运动、对抗运动、与动物运动. Andrej Karpathy 等人[14]在 Sports-1M 数据集上进行预训练,通过所得结果分析出网络能够学习强有力的运动特征,并将这些特征扩展到含有 101 个类别的 UCF101 数据集. Du Tran 和 Jamie Ray 等人在文献[27]中提出了一种搜索时空特征学习的深度三维残差卷积网络,并采用数据集 Sports-1M 进行预训练,在 UCF101 和 HMDB51 数据集上获得了较好的识别效果. Zhaofan Qiu 等人[65]结合残差学习的思想,提出了三种不同的伪三维卷积残差单元,分别使用串行、并行和带捷径(shortcut)的串行三种方式来确定空间卷积和时域卷积之间的关系,在 ImageNet 和 Sports-1M 数据集上共同预训练后,再在目标数据集 UCF101 上进行微调.

Kinetics[63]数据集是 DeepMind 公布的一个人类动作视频数据库. 该数据集包含 400 个人类动作类别,每一个动作类别至少有 400 个视频片段,每个片段持续大约 10 秒,来自于 YouTube 视频. Du Tran 等人[66]在时间分量和空间分量上将三维卷积核进行分解,提出了一种新的时空卷积块“R(2 + 1)D”,并结合二流网络结构,通过在 Sports-1M 数据集下的预训练以及 Kinetics 数据集下的预训练,该方法在目标数据集 UCF101 和 HMDB51 上都获得了不错的识别效果. Saining Xie 等人[67]考虑到用时空可分离的三维卷积替代三维卷积,提出了一个 S3D 结构并在 S3D 的顶部探索了时空特征门,得到 S3D-G 结构,利用 ImageNet 和 Kinetics 数据集共同预训练来初始化该网络结构,最后在 UCF101 和 HMDB51 数据集上微调. Joao Carreira 和 Andrew Zisserman 在文献[68]中提出了双流膨胀的三维卷积神经网络,该网络是基于二维卷积神经网络的一个膨胀,将图像分类中卷积网络的滤波器和池化核扩展为三维的,

并通过 ImageNet 和 Kinetics 数据集共同预训练,从视频中学习时空特征,获得了当下最好的动作识别效果.

I3D-RGB[68]深度模型算法在没有预训练时,它在数据集 Kinetics 上 top1 和 top5 识别准确率分别为 67.5% 和 87.2%,由文献[68]中的数据可以知道,通过数据集 ImageNet 预训练,它的 top1 和 top5 识别准确率分别提升了 4.6% 和 3.1%. R(2+1)D-RGB[66]和 R(2+1)D-Flow[66]深度模型算法在没有预训练时 top1 识别准确率分别为 72% 和 67.5%,top5 识别准确率分别为 90% 和 87.2%,通过 Sports-1M 数据集进行预训练后,它们的 top1 识别率分别提升了 2.3% 和 1%,top5 识别率分别提升 1.4% 和 0.9%. 由此可见,经过预训练后大多数模型都可以获得较大的性能提升.

此外,R(2+1)D-Flow[66]、R(2+1)D-RGB[66]和 R(2+1)D-TwoStream[66]这三个深度模型算法在数据集 Sports-1M 上预训练后在目标数据集 UCF101 上的平均识别准确率分别为 93.3%、93.6% 和 95%,在目标数据集 HMDB51 上平均识别准确率分别为 70.1%、66.6% 和 72.7%. 而它们在数据集 Kinetics 上预训练后在目标数据集 UCF101 上识别准确率分别提升了 2.2%、3.2% 和 2.3%,在目标数据集 HMDB51 上识别准确率分别提升了 6.3%、7.9% 和 6%. 由此可见,同一种算法在不同的数据集上预训练会有不同的性能体现,这可能与预训练数据集的全面性以及与目标数据集的相似性有关联.

5 结论及未来可能的研究方向

视频中的人体动作识别是一个十分重要的研究领域,具有广泛的应用前景. 本文通过三个维度对基于深度模型的动作识别方法进行了综述,并在两个十分流行的动作识别数据集 UCF101 和 HMDB51 上对当下比较经典的动作识别算法进行了对比分析. 结果表明结合传统手工 IDT 特征和深度模型特征方法、多流模型方法、应用大型图像数据集进行预训练都可以获得较好的性能提升. 虽然当前基于深度模型的方法取得了很好的成果,但是依然面临很多的挑战,离实际应用还有距离,在以下几方面有待进一步研究.

(1)视频预处理

对人体动作视频进行预处理后可得到不同输入类型,在这些不同输入类型中,人工设计的光流特征在动作识别上有着广泛而有效的应用. 虽然光流特征带来了不错的识别效果,但是基于光流输入的方法存在一个很大的缺陷,即光流计算的空间和时间成本很高. 鉴于此,设计出更简单高效的时间流输入应该是未来值得探索的一个方向.

(2)视频运动信息表征

目前视频中人体动作识别的研究大多数使用的是基于图像输入并在图像集上预训练好的卷积神经网络模型来学习视频特征,这种方法能够有效学习空间外观信息,但是对于学习长期的运动信息有一定的局限性,即忽略了连续视频帧之间的联系以及视频中的运动信息. 当然也有一些研究者采用长期依赖关系、时空交互等方法来弥补视频转换到图像过程中的信息缺失问题,取得了一定的效果. 但是,除了这种“缺失——再弥补”的方法外,能否一开始就把运动信息或视频帧之间的信息考虑进去? 即设计出针对视频输入的能更好学习运动特征的深度神经网络模型,也是值得研究的一个方向.

(3)模型学习训练

研究者通过改进单个网络模型的结构、增加网络的“深度”、通道的“宽度”、以及融合不同网络模型等方法来不断提升动作识别的性能. 如何探索出网络的“新维度”以及不同网络模型之间的融合训练学习方法有待进一步的研究.

参考文献

[1] 胡琼,秦磊,黄庆. 基于视觉的人体动作识别综述[J]. 计算机学报,2013,36(12):2512-2524.
HU Qiong, QIN Lei, HUANG Qing. Overview of human action recognition based on vision[J]. Chinese Journal of Computers, 2013, 36(12): 2512-2524. (in Chinese)

[2] POPPE R. A survey on vision-based human action recognition[J]. Image and Vision Computing, 2010, 28(6): 976-990.

[3] WEINLAND D, RONFARD R, BOYER E. A survey of vision-based methods for action representation, segmentation and recognition[J]. Computer Vision and Image Understanding, 2011, 115(2): 224-241.

[4] 杜友田,陈峰,徐文立,李永彬. 基于视觉的人的运动识别综述[J]. 电子学报,2007,35(1):84-90.
DU You-tian, CHEN Feng, XU Wen-li, LI Yong-bin. A survey on the vision-based human motion recognition[J]. Acta Electronica Sinica, 2007, 35(1): 84-90. (in Chinese)

[5] CHAQUET J M, CARMONA E J, FERNANDEZ-CABALLERO A. A survey of video datasets for human action and activity recognition[J]. Computer Vision and Image Understanding, 2013, 117(6): 633-659.

[6] DAWN D D, SHAIKH S H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector[J]. Visual Computer, 2016, 32(3): 289-306.

[7] 朱红蕾,朱昶胜,徐志刚. 人体行为识别数据集研究进展[J]. 自动化学报,2018,44(6):978-1004.
ZHU Hong-lei, ZHU Chang-sheng, XU Zhi-gang. Research

progress on human action recognition datasets[J]. Acta Automatica Sinica,2018,44(6):978 – 1004. (in Chinese)

[8] ZHU F,SHAO L,XIE J,et al. From handcrafted to learned representations for human action recognition[J]. Image and Vision Computing,2016,55(P2):42 – 52.

[9] HERATH S,HARANDI M,PORIKLI F. Going deeper into action recognition:a survey[J]. Image and Vision Computing,2017,60(4):4 – 21.

[10] SARGANO A,ANGELOV P,HABIB Z. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition[J]. Applied Sciences,2017,7(1):110 – 147.

[11] WU D,SHARMA N,BLUMENSTEIN M. Recent advances in video-based human action recognition using deep learning:a review[A]. International Joint Conference on Neural Networks[C]. USA:IEEE,2017. 2865 – 2872.

[12] YAO G L,LEI T,ZHONG J D. A review of convolutional-neural-network-based action recognition[J]. Pattern Recognition Letters,2019,118(2):14 – 22.

[13] DU T,BOURDEV L,FERGUS R,et al. Learning spatiotemporal features with 3D convolutional networks[A]. International Conference on Computer Vision[C]. Chile:IEEE,2015. 4489 – 4497.

[14] KARPATHY A,TODERICI G,SHETTY S,et al. Large-scale video classification with convolutional neural networks[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2014. 1725 – 1732.

[15] FEICHTENHOFER C,PINZ A,ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 1933 – 1941.

[16] SIMONYAN K,ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[A]. Neural Information Processing Systems[C]. Canada:NIPS Proceedings,2014. 568 – 576.

[17] WANG H,SCHMID C. Action recognition with improved trajectories[A]. International Conference on Computer Vision[C]. Australia:IEEE,2013. 3551 – 3558.

[18] WANG L,XIONG Y,WANG Z,et al. Temporal segment networks:towards good practices for deep action recognition[J]. ACM Transactions on Information Systems,2016,22(1):20 – 36.

[19] SUN S,KUANG Z,OUYANG W,et al. Optical flow guided feature:a fast and robust motion representation for video action recognition[A]. Computing Vision and pattern Recognition[C]. USA:IEEE,2018. 1390 – 1399.

[20] ZHANG B,WANG L,WANG Z,et al. Real-time action recognition with enhanced motion vector CNNs[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 2718 – 2726.

[21] WANG L,GE L,LI R,et al. Three-stream CNNs for action recognition[J]. Pattern Recognition Letters,2017,92(C):33 – 40.

[22] SHI Y,TIAN Y,WANG Y,et al. Sequential deep trajectory descriptor for action recognition with three-stream CNN[J]. IEEE Transactions on Multimedia,2017,19(7):1510 – 1520.

[23] BILEN H,FERNANDO B,GAVVES E,et al. Dynamic image networks for action recognition[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 3034 – 3042.

[24] FERNANDO B,GAVVES E,ORAMAS M J O,et al. Rank pooling for action recognition[J]. IEEE Transactions on Pattern Analysis Machine Intelligence,2017,39(4):773 – 787.

[25] 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.

[26] 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.

[27] TRAN D,RAY J,SHOU Z,et al. ConvNet architecture search for spatiotemporal feature learning[J]. Computing Research Repository,2017,16(8):1 – 12.

[28] ZHOU Y,SUN X,ZHA Z-J,et al. MiCT:mixed 3D/2D convolutional tube for human action recognition[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2018. 449 – 458.

[29] AHSAN U,SUN C,ESSA I. DiscrimNet:semi-supervised action recognition from videos using generative adversarial networks[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2018. 230 – 240.

[30] GOODALE M A,MILNER A D. Separate visual pathways for perception and action[J]. Trends in Neurosciences,1992,15(1):20 – 25.

[31] 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.

[32] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science,2015,10(4):1 – 4.

[33] SZEGEDY C,LIU W,JIA Y,et al. Going deeper with convolutions[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2015. 1 – 9.

[34] WANG L,XIONG Y,WANG Z,et al. Towards good

practices for very deep two-stream ConvNets[J]. Computer Science,2015,8(7):1 – 5.
[35] FEICHTENHOFER C,PINZ A,WILDES R P. Spatiotemporal residual networks for video action recognition[A]. Neural Information Processing Systems[C]. Spain:NIPS Proceedings,2016. 3468 – 3476.
[36] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 770 – 778.
[37] FEICHTENHOFER C,PINZ A,WILDES R P. Spatiotemporal multiplier networks for video action recognition[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2017. 7445 – 7454.
[38] NG Y H,HAUSKNECHT M,VIJAYANARASIMHAN S,et al. Beyond short snippets:deep networks for video classification[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2015. 4694 – 4702.
[39] SUN L,JIA K,CHEN K,et al. Lattice long short-term memory for human action recognition[A]. International Conference on Computer Vision[C]. Italy:IEEE,2017. 2166 – 2175.
[40] WANG Y,WANG S,TANG J,et al. Hierarchical attention network for action recognition in videos[J]. Computing Research Repository,2016,21(7):41 – 50.
[41] WANG Y,LONG M,WANG J,et al. Spatiotemporal pyramid network for video action recognition[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2017. 2097 – 2106.
[42] VAROL G,LAPTEV I,SCHMID C. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(6):1510 – 1517.
[43] DIBA A,SHARMA V,GOOL L V. Deep temporal linear encoding networks[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2017. 1541 – 1550.
[44] SOOMRO K,ZAMIR A R,SHAH M. UCF101:a dataset of 101 human actions classes from videos in the wild[J]. Computer Science,2012,3(12):2 – 9.
[45] KUEHNE H,JHUANG H,GARROTE E,et al. HMDB:a large video database for human motion recognition[A]. International Conference on Computer Vision[C]. Spain:IEEE,2011. 2556 – 2563.
[46] YANG X,TIAN Y L. Effective 3D action recognition using EigenJoints[J]. Journal of Visual Communication and Image Representation,2014,25(1):2 – 11.
[47] BOBICK A,DAVIS J. An appearance-based representation of action[A]. International Conference on Pattern Recognition[C]. Austria:IEEE,1996. 307 – 312.
[48] WEINLAND D,RONFARD R,BOYER E. Free viewpoint action recognition using motion history volumes[J]. Computer Vision and Image Understanding,2006,104(2):249 – 257.
[49] BOBICK A F,DAVIS J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2001,23(3):257 – 267.
[50] YILMAZ A,SHAH M. Actions sketch:a novel action representation[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2005. 984 – 989.
[51] LINDEBERG T,LAPTEV I. On space-time interest points[J]. International Journal of Computer Vision,2005,64(2 – 3):107 – 123.
[52] HARRIS C. A combined corner and edge detector[A]. Alvey Vision Conference[C]. UK:IEEE,1988. 1 – 6.
[53] WILLEMS G,TUYTELAARS T,GOOL L. An efficient dense and scale-invariant spatio-temporal interest point detector[A]. European Conference on Computer Vision[C]. France:Springer,2008. 650 – 663.
[54] DALAL N,TRIGGS B. Histograms of oriented gradients for human detection[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2005. 886 – 893.
[55] KLASER A. A spatiotemporal descriptor based on 3D-gradients[A]. British Machine Vision Conference[C]. LEEDS:BMVA,2008. 1 – 10
[56] LAPTEV I,MARSZALEK M,SCHMID C,et al. Learning realistic human actions from movies[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2008. 24 – 32.
[57] DALAL N,TRIGGS B,SCHMID C. Human detection using oriented histograms of flow and appearance[A]. European Conference on Computer Vision[C]. Austria:Springer,2006. 428 – 441.
[58] OJALA T,PIETIK,INEN M,et al. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(7):971 – 987.
[59] 田国会,尹建芹,闫云章,李国栋. 基于混合高斯模型和主成分分析的轨迹分析行为识别方法[J]. 电子学报,2016,44(1):143 – 149.
TIAN Guo-hui,YIN Jian-qin,YAN Yun-zhang,LI Guo-dong. Gaussian mixture models and principal component analysis based human trajectory behavior recognition[J]. Acta Electronica Sinica,2016,44(1):143 – 149. (in Chinese)
[60] KLASER A,SCHMID C. Action recognition by dense trajectories[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2011. 3169 – 3176.
[61] FAN L,HUANG W,GAN C,et al. End-to-end learning of

motion representation for video understanding[A]. Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 6016 – 6025.

[62] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[A]. Computer Vision and Pattern Recognition[C]. USA: IEEE, 2009. 248 – 255.

[63] KAY W, CARREIRA J, SIMONYAN K, et al. The Kinetics human action video dataset[J]. Computing Research Repository, 2017, 19(5): 50 – 72.

[64] SUN L, JIA K, YEUNG D Y, et al. Human action recognition using factorized spatio-temporal convolutional networks[A]. International Conference on Computer Vision[C]. Chile: IEEE, 2015. 4597 – 4605.

[65] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[A]. International Conference on Computer Vision[C]. Italy: IEEE, 2017. 5534 – 5542.

[66] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[A]. Computer Vision and Pattern Recogmtion[C]. USA: IEEE, 2018. 6450 – 6459.

[67] XIE S, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning for video understanding[J]. Computing Research Repository, 2018, 27(7): 1 – 10.

[68] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the Kinetics dataset[A]. Computer Vision and Pattern Recognition[C]. USA: IEEE, 2017. 4724 – 4733.

作者简介

罗会兰　女,1974 年 9 月生于江西上高. 2008 年获浙江大学工学博士学位. 现为江西理工大学图像处理实验室教授、硕士生导师. 主要从事机器学习、模式识别等方面的研究.
E-mail: luohuilan@ sina. com

童　康　男,1992 年 2 月生于江苏南京. 2016 年进入江西理工大学,在读硕士研究生. 研究方向为视频动作识别.
E-mail: 2804947614@ qq. com