计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190300005

基于深度学习的人体行为识别方法综述

蔡 强[1,2] 邓毅彪[1,2] 李海生[1,2] 余 乐[1,2] 明少锋[1]
1 北京工商大学计算机与信息工程学院 北京 100048
2 食品安全大数据技术北京市重点实验室 北京 100048
(caiq@th.btbu.edu.cn)

摘 要 人体行为识别作为计算机视觉领域的重要研究热点,在智能监控、智能家居、虚拟现实等诸多领域中具有重要的研究意义和广泛的应用前景,备受国内外学者的关注。基于传统手工特征的方法难以处理复杂场景下的人体行为识别。随着深度学习在图像分类方面取得巨大成功,将深度学习用于人体行为识别方法中已逐渐成为一种发展趋势,但其仍然存在一些困难与挑战。首先,根据特征提取方法的不同,简单回顾了早期基于传统手工特征的行为识别方法;然后,从网络结构的角度着重对近年来一些基于深度学习的人体行为识别方法进行论述和分析,其中包括目前常用的双流网络架构和三维卷积网络架构等;另外,还介绍了目前用于评价方法性能的人体行为识别数据集,同时总结了部分典型方法在 UCF-101 和 HMDB51 两个著名的公开数据集上的性能;最后,从性能和应用两个方面对基于深度学习的人体行为识别方法的未来发展方向进行了展望,并指出了当前方法存在的不足之处。

关键词:人体行为识别;深度学习;卷积神经网络;人体行为识别数据集
中图法分类号 TP391

Survey on Human Action Recognition Based on Deep Learning

CAI Qiang[1,2],DENG Yi-biao[1,2],LI Hai-sheng[1,2],YU Le[1,2] and MING Shao-feng[1]
1 School of Computer and Information Engineering,Beijing Technology and Business University,Beijing 100048,China
2 Beijing Key Laboratory of Big Data Technology for Food Safety,Beijing 100048,China

Abstract As an important research hotspot in the computer vision community,human action recognition has important research significance and broad application prospects in many fields such as intelligent surveillance,smart home and virtual reality,and it has attracted the attention of scholars at home and abroad. For the methods based on traditional handcrafted features,it is difficult to deal with human action recognition in complex scenarios. With the great successes of deep learning in image classification,the application of deep learning to human action recognition has gradually become a development trend,but there are still some difficulties and challenges. In this paper,firstly,according to the difference of the feature extraction approaches,the early traditional handcrafted representation-based methods for human action recognition were simply overviewed. Then,from the perspective of network architecture,some deep learning-based approaches for human action recognition were discussed and analyzed,including Two-Stream Networks,3D Convolutional Networks,etc. Besides,this paper introduced the current human action recognition datasets used to evaluate the methods performance,and summarized the performance of some typical methods on two well-known public datasets of UCF-101 and HMDB-51. Finally,the future trends of deep learning-based methods were discussed from two aspects of performance and application,and the shortcomings were also pointed out.

Keywords Human action recognition,Deep learning,Convolutional neural network,Human action recognition dataset

1 引言

近年来,人体行为识别作为计算机视觉领域的一项重要研究内容,受到了国内外学者的广泛关注和研究。相比于静态图像中的人体识别研究,视频人体行为识别更关注人体在视频图像序列中的时空变化,其根据人体行为视频中的图像

到稿日期:2019-03-06 返修日期:2019-07-23 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61877002);国家社会科学基金项目(18BJL202);教育部人文社会科学基金项目(17YJCZH127);北京市教委项目(PXM2019_014213_000007);北京市科技计划(Z161100_001616004)
This work was supported by the National Natural Science Foundation of China(61877002),National Social Science Fund of China(18BJL202),MOE (Ministry of Education in China) Project of Humanities and Social Sciences (17YJCZH127),Beijing Municipal Commission of Education,China(PXM2019 014213 000007) and Science and Technology Program of Beijing,China(Z161100_001616004).
通信作者:邓毅彪(dyb9714@sina.com)

帧或图像序列，通过计算机对视觉信息的处理和分析，自动识别出人体做出的行为动作[1]。作为视频理解中的关键技术，视频人体行为识别在智能监控、智能家居、虚拟现实、视频检索等众多场景中具有重要的研究意义和广泛的应用前景。人体行为的多样性，以及视频的视角变化和非刚性运动的复杂性，使得人体行为识别存在巨大的挑战。图 1 为人体行为识别的流程图。

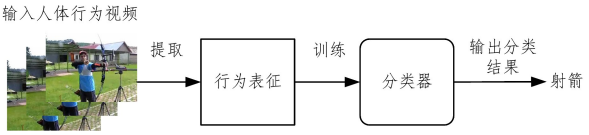

图 1 人体行为识别流程

Fig. 1 Pipeline of human action recognition

人体行为特征是从视频数据中提取到的关键信息的表征，是行为识别中的关键部分，其好坏直接影响识别的性能。人体行为的分类是指将人体行为特征向量作为输入，通过机器学习的方法训练一个分类器，将需要识别的特征向量输入该分类器中，从而得到类别的分类结果。

在过去的几十年里，国内外学者们提出了大量的视频行为识别方法，并给出了一些公开的行为识别数据集。卷积神经网络(Convolutional Neural Network，CNN)[2]自 2012 年被提出并广泛应用于图像领域以来，就以其卓越的性能在图像识别、目标检测、场景分类等领域取得显著成果，由此 CNN 从图像领域扩展到了视频领域并取得了重大进展。

目前，已有较多学者从不同角度对行为识别方法进行了总结和综述。Aggarwal 等[3]按复杂程度把行为识别分为 4 类：姿势(Gestures)、个体动作(Actions)、交互动作(Interactions)以及团体活动(Group Activities)。其中，姿势是指人体肢体部分的基础移动，如摇头、挥手等，是行为分类中复杂度最低的行为；个体动作是指单个人的基本运动动作，如跑步、跳高等，也可以看作多个 Gestures 的组合；交互动作是人人交互和人物交互的合称，如打斗、握手、弹钢琴等，也是目前最受关注的行为识别类型；团体活动是指一个场景中包含多人和多物的活动，如一群人的比赛、团体会议等，也是最复杂的行为识别类型。Hassner[4]对早期的行为识别数据集和基准进行了详细的回顾，并总结了 2013 年以前的行为识别系统需要克服的多种挑战；Huang 等[1]对智能视频监控技术的发展、现状以及典型算法给出了较为全面的综述，以底层、中层、高层的方式对智能视频监控技术流程进行划分，并对目标检测、跟踪、行为分类识别以及行为分析算法进行了总结；Zhu 等[5]首次对手工行为表征和基于学习的行为表征做出了详细的综述，并讨论了现有的两类方法中存在的限制和优势；Luo 等[6]着重综述了基于传统手工特征的行为识别方法，并对基于深度学习的方法做出了简单的回顾。

目前，视频人体行为识别的方法按特征提取方式的不同分为两类。1)基于传统手工特征的行为识别方法。其首先利用专家设计的特征提取符提取视频的底层行为特征；之后通常采用主成分分析(Principal Component Analysis，PCA)和白化(Whitening)对底层特征进行预处理，以消除数据间的相关性，防止过拟合；然后将特征编码成定长的特征向量，并将其作为行为分类器的输入进行训练，最终得到训练好的行为分类器。常见的编码方式有矢量量化(Vector Quantization)、软分配(Soft-Assignment)、稀疏编码(Sparse Coding)、费舍尔矢量(Fisher Vector)，分类器通常采用支持向量机(Support Vector Machine，SVM)。2)基于深度学习的方法。其利用迭代学习自动从视频中提取深度学习的行为特征向量，然后通过深度模型得到类别得分，并根据数据的标签，利用反向传播的方式调整网络模型参数，最终达到良好的分类效果。模型最后的全连接层和 Softmax 层相当于分类器，整个模型的训练是一个端到端学习的过程。图 2 给出了目前人体行为识别的分类情况。

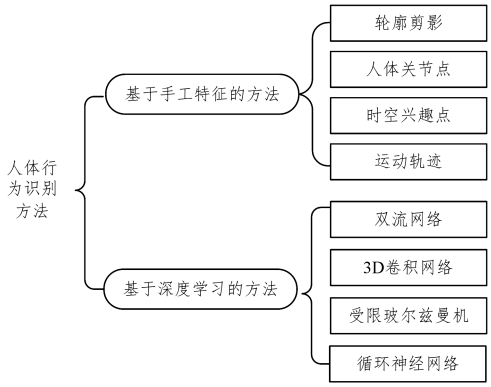

图 2 人体行为识别方法的分类

Fig. 2 Categorization of human action recognition approaches

前人的研究论述中没有包括目前最先进的方法，为了让初学者对视频行为识别方法有更好的理解，更快地掌握视频行为识别发展的前沿动态，本文分别综述了基于传统手工特征的行为识别方法和基于深度学习的方法，并重点分析了当前性能优越的深度学习模型。本文主要贡献如下：

1)简要介绍了现有的基于传统手工特征的行为识别方法，将其按人体行为特征提取方法的特点分类，并总结了典型方法的优缺点；

2)重点阐述了基于深度学习的方法，将其按结构特点分类，并详细地分析了其中典型的方法；

3)对目前国内外公开的行为识别数据集进行介绍和分析，并在常用的数据集上对部分方法进行了比较；

4)对目前视频人体行为识别方法存在的挑战和问题展开讨论，并展望了未来的发展方向。

2 基于手工特征的人体行为识别方法

基于手工特征的人体行为识别方法的流程通常如图 3 所示。首先，对视频数据的连续帧进行采样，得到采样点；然后，按照专家设计的手工特征提取方法对采样点提取手工特征；其次，将提取到的手工特征编码成特征向量；接着，将特征向量输入到行为分类器中进行训练；最后，将对测试视频提取的手工特征向量输入训练好的分类器，得到分类结果。在深度学习未被引用到人体行为识别领域之前，国内外学者设计了多种手工特征，如轮廓剪影(Human Silhouette)、人体关节点(Human Joint Point)、时空兴趣点(Space-Time Interest Points)、运动轨迹(Trajectories)等，并进行了大量尝试。

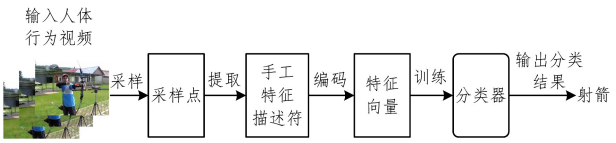

图 3　基于手工特征的行为识别方法流程

Fig. 3　Pipeline of human action recognition approach based on handcrafted features

2.1　基于轮廓剪影的特征提取方法

早期的工作将视频帧看作整体，提取全局特征。通常采用背景剪除法、人体轮廓剪影、差分法等检测人体行为区域，然后对整个区域提取特征作为行为表征。Bobick 等[7]提出利用背景剪除法从视频中获得前景的轮廓特征，然后将所有图像帧的轮廓特征叠加以获取帧间差，建立具有运动效果的运动能量图(Motion Energy Image，MEI)，并根据轮廓运动的时间函数构造运动历史图(Motion History Image，MHI)。这种方法描述能力强，包含信息多，在简单背景下容易提取感兴趣区域；但是对噪声、视角和相互遮挡特别敏感，在复杂背景下难以得到人体运动区域的信息，难以提取轮廓特征，存在较大的局限性。

2.2　基于人体关节点的特征提取方法

基于人体关节点的特征提取方法通过姿态估计获取人体各个关节点的位置以及关节点位置的运动信息，从而对人体行为进行表征。Fujiyoshi 等[8]使用头部加四肢的 5 个关节点组成的星形图来表示当前视频中人体的姿态，并将行为的特征向量表示为 5 个关节点与重心构成的矢量。这种采样方法通常需要先对视频前景和背景进行分割，然后利用运动检测算法对视频中的人体定位。随着深度相机的出现和普及，研究者们开始利用深度图像提取人体关节点位置信息，简化了基于人体关节点的特征提取方法。Yang 等[9]从人体行为的深度图像中采集人体关节点的三维坐标，将这些关节点形成的人体轮廓作为行为特征用于识别，取得了不错的效果。通过这种方法提取到的行为特征信息较为完整，但是该方法受限于深度图像不便于获取以及人体几何的非刚性建模难以用简单的数学模型描述的问题。

2.3　基于时空兴趣点的采样方法

基于时空兴趣点的采样方法通过探测算子检测出视频的时空兴趣点，然后从兴趣点周围提取行为特征。Laptev[10]将 Harris 角点[11]的二维空间兴趣点扩展到三维时空兴趣点，对行为视频检测 Harris3D 兴趣点。Harris3D 检测在时空上具有显著变化的区域，并且能够自适应地选择兴趣点的时空范围，最后通过统计像素直方图形成描述行为的特征向量。但是，Dollar 等[12]指出这种方法检测到的稳定兴趣点数量不足，因此提出在时间维度上采用 Gabor 滤波器并在空间维度上采用高斯滤波器进行滤波，由此检测到的兴趣点数量将随局部邻域块尺寸的变化而变化。接着，Willems 等[13]提出基于 Harris3D 的时空兴趣点检测方法，该方法在多尺度的视频三维数据中通过对 Hessian 矩阵的计算筛选出兴趣点所在的时空位置，大幅降低了兴趣点检测的时间复杂度。然而，这样检测到的兴趣点依然过于稀疏。Wang 等[14]提出稠密网格的方法在视频数据上提取行为特征，并将其与多种局部特征描述符如梯度直方图(Histogram of Oriented Gradient，HOG)[15]、光流梯度直方图(Histograms of Oriented Optical Flow，HOF)[16]和运动边界直方图(Motion of Boundary History，MBH)[17]进行比较，实验结果表明稠密的采样方式优于稀疏的兴趣点采样方式。

基于时空兴趣点的特征提取方法不需要对视频的前景和后景做分割处理，在背景相对复杂的情况下效果较好，对遮挡和光照的影响比较敏感。基于时空兴趣点的特征提取方法不需要对视频的前景和后景做分割处理，在背景相对复杂的情况下效果较好，对遮挡和光照的影响比较敏感。

2.4　基于轨迹跟踪的特征提取方法

基于轨迹跟踪的特征提取方法通过追踪人体动作的运动轨迹来提取行为特征。运动轨迹包含明确的结构信息，加强了采样点之间的联系。近年来，基于轨迹的行为表征被证明是最成功的手工浅层表征。Wang 等[18]提出的稠密轨迹(Dense Trajectories)方法利用光流场获取视频帧序列的密集采样点的运动轨迹，再沿着轨迹提取 HOG[15]，HOF[16]，MBH[17]特征，其流程如图 4 所示。

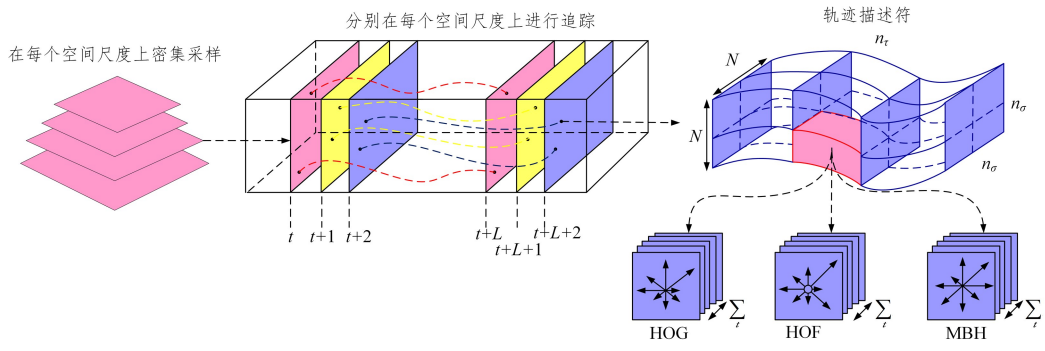

图 4　稠密轨迹的方法

Fig. 4　Illustration of dense trajectory description

这种方法能够排除背景的干扰，获取更多运动信息，对视频的表征能力较强；但是提取光流场的计算开销大，效率较低，且相机本身的运动会造成较大干扰。接着，WANG 等进一步引入了一个整合了 HOG[15]，HOF[16]，MBH[17]的特征，提出了改进的稠密轨迹方法(improved Dense Trajectories，iDT)[19]。该方法对轨迹施加全局平滑约束，得到更鲁棒的运动轨迹，并利用 SURF(Speed Up Robust Features，SURF)[20]匹配不同帧间的特征点来估计相机运动，以区分人体运动和

摄像机引起的运动，提高了特征的鲁棒性。在深度学习被广泛应用于人体行为识别领域之前，iDT 算法是基于传统手工特征的方法中效果最好、最鲁棒的，许多基于深度学习的方法在与 iDT 算法结合后取得了较大的性能提升；但是，由于训练阶段和测试阶段都无法避免高计算复杂度，因此该算法的速度较慢。

3 基于深度学习的人体行为识别方法

对于不同复杂场景中存在的光照、遮挡、视角变化等问题，传统手工特征并不具有普适性，因此以深度学习的方式从数据中自动学习特征可能更有效。与基于传统手工特征的方法不同，基于深度学习的人体行为识别方法的流程框架如图 5 所示。

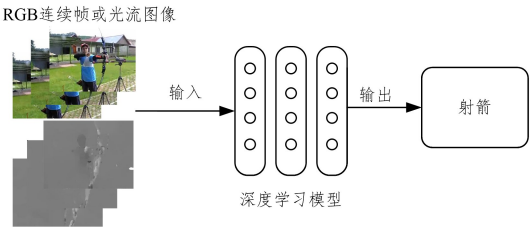

图 5 基于深度学习的人体行为识别方法

Fig. 5 Human action recognition approach based on deep learning

基于深度学习的人体行为识别方法以端到端的方式，利用可训练的特征提取模型从视频中自动学习行为表征来完成分类。目前，基于深度学习的行为识别方法的网络结构主要有双流网络（Two-Stream Network）和 3D 卷积网络（3D Convolutional Network）两种；还有部分学者提出了其他思路，如受限玻尔兹曼机（Restricted Boltzmann Machines，RBM）、循环神经网络（Recurrent Neural Networks，RNN）、独立子空间分析（Independent Subspace Analysis，ISA）等，它们也取得了不错的效果。

3.1 双流网络

基于双流网络结构的行为识别方法自 2014 年由 Simonyan[21] 提出后，就得到了行为识别研究者的广泛关注，其基本流程如图 6 所示。双流网络结构分为时间流卷积神经网络和空间流卷积神经网络两个分支，且两个分支具有相同的网络结构。时间流卷积神经网络先对视频序列中相邻两帧计算光流图像，再对多帧堆叠的光流图像提取时序（Temporal）信息；空间流卷积神经网络则对视频 RGB 图像提取空间（Spatial）特征，最后将两个网络分别得到的得分融合，从而得到最终的分类结果，这种方法大幅提高了视频行为识别的准确率。

Wang 等[22-23] 在此基础上提出时域分割网络（Temporal Segment Network，TSN），采用稀疏时间采样策略将视频进行时域分割后随机抽取片段，以解决双流网络对时间较长的视频的建模能力不足的问题。

Feichtenhofer 等[24] 发现了双流架构的两个问题：1）不能在空间和时间特征之间学习像素级的对应关系；2）空域卷积只在单 RGB 帧上，时域卷积只在堆叠的 L 个时序相邻的光流帧上，时间规模非常有限，导致不能利用视频中的两个非常重要的线索来完成行为识别，即无法在指定的外观位置区域（空间线索）同时观察对应的光流（时间线索）有何变化，以及空间线索随时间的变化趋势。

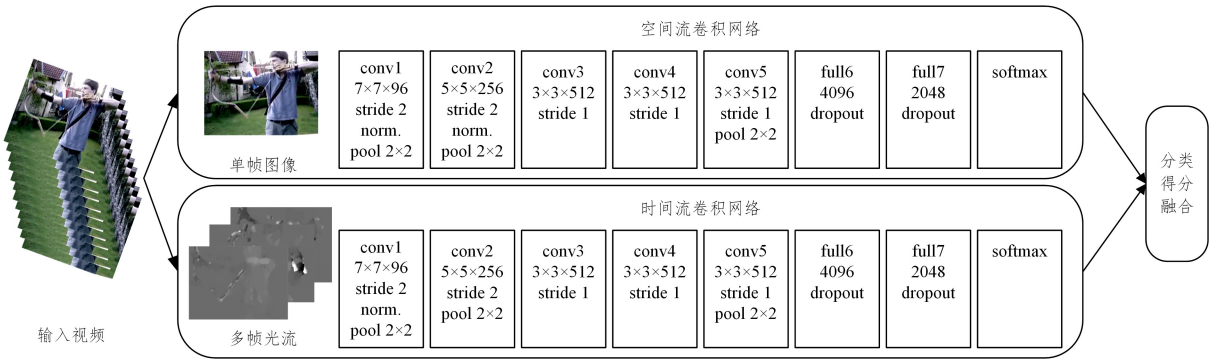

图 6 双流网络框架

Fig. 6 Architecture of two-stream network

受残差网络[25] 在图像识别领域中表现出色的启发，Feichtenhofer 等[26] 在双流网络的基础上提出了时空残差网络（Spatiotemporal Residual Networks，STResNet，STResNet）。STResNet 在时域上进行了扩展，并延续了网络融合的特色，找到了时域和空域的像素级对应关系。在双流网络中，时域网络和空域网络也利用残差连接（Residual Connection）进行参数的传递，在处理视频数据过程中，这样的时空连贯性是一个重要的线索。另外，作者采用了新的时间残差的方法，使用了小的非对称滤波器，通过将残差路径中的空间滤波器维数映射转换为时间滤波器进行时间卷积。该时间滤波器可以学习场景与运动的变化。通过堆叠这些滤波器，可以学习到更深层次的时空特征。

Fernando 等[27-28] 从基于视频帧的 CNN 特征中有效地编码视频片段的时序信息，并利用 Rank-Pooling 的池化方法，通过双层优化方式端到端地学习模型的所有参数和视频表征。Bilen 等[29] 在此基础上提出动态图网络（Dynamic Image Networks），即将视频汇集成一张称作动态图（Dynamic Image，DI）的图像，然后通过学习一个排名机（Ranking Machine）来捕捉数据的时序演变，并将排名机的参数作为一个表征。作者利用双流网络的思想，将静态图像、动态图像、光流图像、动态光流图像作为输入，提出了一个四流网络模型，将其与 iDT 特征结合，取得了不错的识别效果。这种方法的特点是将视频转换成一张 DI，使得现有的 CNN 模型对静止图像预训练后能立即扩展到视频上。

Gao 等[30]利用大量的无标签视频数据训练 Im2Flow 网络，让网络模型从大量的无标签行为数据中学习行为先验，使模型能够通过行为先验生成预测光流图像；然后向模型输入一张静态图像，得到预测的光流图像，将静态图像和预测得到的光流图像作为双流网络的输入。这种无监督学习的方式适用于只有少量标签的行为识别数据。

Wang 等[31]提出了三流卷积网络架构，该架构在双流网络的基础上将时间流进一步分为局部时间流和全局时间流，然后将 RGB 特征和 Flow 特征分别作为空间流和局部时间流的输入，并将学习运动叠差图像（Motion Stacked Difference Image，MSDI）的 CNN 特征作为全局时间流的输入。三流卷积网络相比于双流网络[21]性能略有提高。

Girdhar 等[32]提出了一种 VLAD 的池化方法用于时空视频特征的融合，并基于双流网络[21]探索了如何聚合视频帧级别的特征以表示整个视频特征，以及如何聚合双流网络的不同流网络提取的特征信息。与常用的最大池化和平均池化相比，ActionVLAD 池化方式可以聚合不同子类特征的描述符来共同描述整个视频的特征，相比文献[21]的方法取得了巨大的性能提升。

Lin 等[33]从行为类别粒度的角度出发，基于双流网络提出了一个粗到细网络（Coarse-to-Fine Networks），从不同的行为类别粒度中提取共享的深度特征，并且逐步整合特征来获得对输入行为更准确的特征表达。接着，通过在不同时间点对不同流的信息进行异步融合，能够更好地利用不同流之间的信息互补。实验结果证明了该方法的先进性。

基于双流网络架构的方法虽然准确率高，但需要预先对视频提取光流图像，并且分开训练两个网络，这个过程非常耗时，难以达到实时性的要求。

3.2　3D 卷积网络

目前，行为识别的方法大多使用基于图像的 2D 卷积神经网络来学习单帧图像的 CNN 特征，往往容易忽略连续帧间的联系，造成视频动作信息的丢失，因此利用 3D 卷积网络来学习视频行为表征成为了行为识别的一个重要研究方向。

基于 3D 卷积网络结构的行为识别方法由 Ji 等[34]在 2010 年首次提出，利用 3D 卷积核进行 3D 卷积，对视频沿着空间和时间维度直接提取时空特征。3D 卷积操作如图 7 所示。

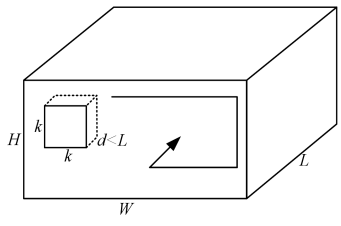

图 7　3D 卷积操作

Fig. 7　Operation of 3D convolution

3D 卷积网络结构由 1 个硬连线层、3 个卷积层、2 个下采样层以及 1 个全连接层组成。硬连线层产生灰度、梯度和光流 3 个通道，然后在每个通道进行卷积和下采样操作，最后连接所有通道的信息计算得到最终的行为表征。随后，Ji 等[35]通过对高层特征进行规则化，将 3D 卷积网络与不同架构相结合。

Tran 等[36]提出了 C3D 网络，通过系统化的研究找到了 3D 卷积最合适的时序卷积核长度，包括 8 层具有 3×3×3 卷积核的卷积层、5 层池化层以及 2 层全连接层。实验证明，C3D 提取的特征是通用、有效且紧凑的。C3D 在 UCF-101[37]数据集上的准确率达到了 82.3%，在 HMDB-51[38]数据集上的准确率达到了 51.6%。随后，Tran 等[39]将残差网络（ResNet）[25]与 C3D 网络相结合，提出了 Res3D 网络。Res3D 网络进一步提升了网络的性能，运行速度比 C3D 快了两倍，模型尺寸是 C3D 的一半，在 UCF-101 和 HMDB-51 数据集上分别提升了 3.5%和 3.3%的准确率。不同于 C3D 和 Res3D 在相对短视频片段中学习时空信息，Varol 等[40]针对人体行为深度卷积特征提取算法难以在持续时间较长的完整人体行为视频中对人体行为建模的问题，提出了一个长时序卷积结构（Long-term Temporal Convolutions，LTC），在较长的视频片段中使用 3D 卷积来表示全时间尺度下的行为，并展示了高质量光流作为 LTC 输入的重要性。由于训练一个 3D 卷积网络的计算和存储开销非常大，为了减少网络参数，降低 3D 卷积核的复杂性和训练样本数据不足带来的影响，Sun 等[41]提出了一个分解时空网络（Factorized Spatiotemporal Convolutional Network，FstCN），将原始的 3D 卷积核分解成在低层学习的一系列 2D 空间卷积核和上层的一个 1D 时间卷积核，并且提出了一种有效的训练和推理策略，对给定的视频序列的多个视频片段采样，解决了序列对齐问题。

为了避免从零训练 3D 卷积网络并且能够使用 2D 卷积学习到的知识，Mansimov 等[42]提出了几种利用 2D 卷积权值初始化 3D 卷积权值的方法，包括平均、缩放、0 权值和负权值初始化。实验结果证明，负权值初始化在所有初始化方法中取得了最好的结果。

通过把一个 3×3×3 的 3D 卷积滤波器分解成一个 1×3×3 的卷积滤波器和一个 3×1×1 的卷积滤波器，Qiu 等[43]基于残差网络设计了 3 种 Bottleneck Building Block，并提出了一个伪 3D 残差网络（Pseudo-3D ResNet）。Pseudo-3D ResNet 不仅显著减小了模型尺寸，而且能够利用在图像数据集中训练好的 2D 卷积神经网络。

通过吸收之前优秀方法的思想，Carreira 等[44]提出了 I3D 网络（Inflated 3D Convolutional Network）。I3D 将用于图像分类的 2D 卷积网络变形成可以提取时空特征的特征提取器，同时利用图像分类任务中预训练好的权值有效进行权值初始化，弥补了 3D 卷积网络参数多以及需要从零开始训练的不足。另外，虽然 3D 卷积已经可以直接从连续的 RGB 图像帧中提取到时空特征，但与双流网络的思想相结合，利用光流图像能够进一步提升方法的识别性能。经过 Kinetics[45]行为识别数据集预训练的双流 I3D 模型，在 UCF-101 数据集上取得了 97.9%的识别准确率，在 HMDB-51 数据集上取得了 80.2%的识别准确率。

Wang 等[46]提出了一种外观关系网络（Appearance-and-Relation Networks，ARTNet），其中包含一种称为 SMART 的通用块。SMART 块由外观和关系两个分支组成，可以同时

对 RGB 输入的外观和关系分开显式建模。外观分支基于每帧像素的线性连接来对视频行为的空间结构建模，而关系分支是基于多帧像素间的乘法运算来捕捉动作的时序动态信息。ARTNet 通过多个 SMART 块堆叠，可以从不同尺度对视频动作的外观和关系建模，这也允许以端到端的方式优化 SMART 块的参数。ARTNet 与 TSN[23] 框架的结合在 UCF-101 数据集上可以达到 94.3% 的识别准确率，在 HMDB-51 数据集上可以达到 70.9% 的识别准确率。

3D 卷积网络对视频连续帧组成的立方体提取 3D 卷积特征，同时捕捉时间维度和空间维度的特征信息，且一次处理多帧图像，加快了运行速度；尽管不需要预先提取光流图像，但是 3D 卷积的计算开销较大，对硬件性能要求更高，并且识别准确率比利用光流图像的双流网络方法略低。因此，目前最先进的方法也结合了双流网络的思想，利用光流图像来进一步提升行为识别方法的性能。

3.3 受限玻尔兹曼机

受限玻尔兹曼机[47]是一种可通过输入数据集学习概率分布的生成网络模型，输入/输出层神经元和隐藏层神经元之间通过一个权值矩阵 $\boldsymbol{w}$ 和偏置向量连接，同一层神经元之间相互独立，不同层之间的神经元相互连接。Taylor 等[48]利用门控受限玻尔兹曼机，以无监督的方式学习视频中的运动信息，并结合卷积来微调网络参数，能够有效地提取运动敏感特征，在 KTH[49] 数据集和 Hollywood2[50] 数据集上取得了不错的效果。Tran[51] 等提出一种利用高斯受限玻尔兹曼机来学习视频中人体运动差异特征的有效方法，定义了一个表示两帧间差异的减函数，从而为动作创建简单的时空显著图，通过消除与行为识别不相关的共有形状和背景图像来突出空间上的运动模式，使浅层的 RBM 能够更容易地学习这些显著图中的动作。

基于受限玻尔兹曼机的方法的最大特点是可以利用无标签数据进行无监督学习，得到时空特征表示方法。

3.4 循环神经网络

循环神经网络[52]常被用于对时间序列数据进行建模，通过控制当前信息和历史信息的贡献度来实现序列信息的累积。因此，RNN 在对时域动态特征建模以及特征学习方面具有强大的性能。然而，RNN 普遍存在梯度消失问题，利用长短时记忆(Long Short-Term Memory，LSTM)[53]可以在一定程度上减轻该问题。LSTM 的结构如图 8 所示，LSTM 单元接收上一时刻的输出隐藏状态 h_{t-1} 和当前输入 X_t，通过输入门 i_t、遗忘门 f_t 以及输出门 o_t 来更新状态，并将当前结果输出。其中，遗忘门决定上一时刻的信息是否需要遗忘；输入门决定当前时刻的信息是否需要保留；输出门用于控制有多少信息从记忆单元传递到下一时刻的隐藏状态 h_t。

Baccouche 等[54]首先将卷积神经网络扩展到 3D 形式，自动学习视频的时空特征，然后训练 LSTM RNN 学习卷积特征描述间的时序信息，最终得到人体行为特征描述。

Donahue 等[55]提出长时循环卷积神经网络(Longterm Recurrent Convolutional Network，LRCN)，并将 CNN 与 LSTM 相结合，通过 CNN 提取单帧图像的卷积特征，将其按时间顺序输入 LSTM 中，最终得到视频数据的行为特征。LRCN 在 UCF-101 数据集上取得了 82.92% 的识别准确率。Wu 等[56]提出的结合双流网络和 LSTM 的混合学习框架，能够对静态空间信息、短时运动、长时时序线索等多个方面建模，将时间流和空间流提取的卷积特征作为 LSTM 网络的输入，由此对长时时序建模。该混合框架在 UCF-101 数据集上取得了 90.1% 的识别准确率。

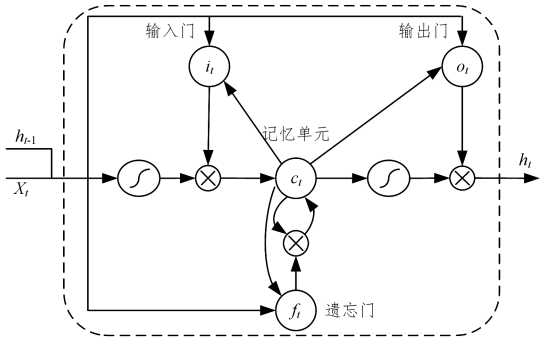

图 8 LSTM 单元的结构

Fig. 8 Structure of LSTM unit

为了更好地利用视频的空间相关性，Li 等[57]提出 VideoLSTM。VideoLSTM 在 LSTM 中加入了基于运动的注意力机制，通过卷积神经网络提取视频图像帧中的空间相关性，并产生基于运动的注意力映射，依靠视频级别的动作标签便可以利用注意力映射来定位动作的时空位置。其在 UCF-101 数据集上取得了 88.9% 的识别准确率，在 HMDB-51 数据集上取得了 56.4% 的识别准确率，且与 iDT 特征结合后识别准确率可以得到进一步提升。

基于循环神经网络的方法能够很好地利用相邻帧间的时间相关性来对视频中人体行为的时序特征建模，但其识别准确率还有待提高。

4 人体行为识别数据集

对于人体行为识别方法的性能，需要在同样的数据集下进行分析比较。目前众所周知的人体行为识别公开数据集的相关信息如表 1 所列。

表 1 人体行为识别数据集

Table 1 Human action reognition datasets

数据集名称	发布时间/年	片段数量/个	类数/类	视频来源
KTH[49]	2004	600	6	志愿者拍摄
Weizmann[58]	2005	90	10	志愿者拍摄
UCF-Sports[59]	2008	150	10	体育电视节目
Hollywood2[50]	2009	3669	12	Hollywood 电影
HMDB-51[35]	2011	6849	51	YouTube 和电影
UCF-101[37]	2012	13320	101	YouTube
Kinetics[45]	2017	306245	400	YouTube

随着人体行为识别受到越来越多学者的关注，国内外公开的行为识别数据集的规模也在逐渐扩大，行为的类型亦越来越多样化，为人体行为识别方法未来的发展提供了可靠的基础。

KTH 数据集在 2004 年公开发布，包括 6 类人体行为：走路，慢跑，跑步，拳击，鼓掌，挥手；由 25 个志愿者在 4 个不同场景下拍摄，包含 600 个人体行为视频序列。但是，由于相机拍摄的角度是固定的，且背景也是静态的，因此只在早期用来评价一些简单的行为识别方法。

Weizmann 数据集发布于 2005 年，由 9 个志愿者参与拍摄，每个志愿者展示 10 个动作，共 90 个视频片段。拍摄背景单一，视角固定，动作简单。

UCF-Sports 数据集包括 10 个体育运动动作，共 150 个视频片段，数据来源于体育电视节目；视频数据的背景多样，且摄影机不固定，具有类内变化。

Hollywood2 数据集包括 12 类人体行为，共 3669 个视频序列，数据均来源于 Hollywood 电影，且电影场景中的视角变换，背景复杂多样。

布朗大学于 2011 年发布的 HMDB-51 数据集包括 51 类人体行为，共 6849 个视频片段，分辨率为 320×240，数据来源于电影和 YouTube 网络视视频库；视频中的背景复杂，行为多样。

2012 年中佛罗里达大学计算机视觉研究中心发布的 UCF-101 数据集包含 101 类人体行为，共 13320 个视频片段，分辨率为 320×240，数据均来源于 YouTube 网络视频库。该数据集为行为识别提供了更丰富的多样性，视频中的人体行为更贴合真实场景，在人体外形和姿态、视觉的角度和焦点、背景以及光照等方面均有较大变化，是目前视频人体行为识别领域最具挑战性的数据集之一，旨在通过学习现实场景中的人体行为来推动视频人体行为识别的发展。

Kinetics 数据集发布于 2017 年，包含 400 类人体行为，共 30 万个视频片段，数据均来源于 YouTube 网络视频库，是目前公开的最大的人体行为识别数据集。

表 2 对比了一些典型方法在两个公开数据集上的性能。其中，卷积网络的输入 OF 表示光流图像(Optical Flow)。

表 2　人体行为识别方法的比较

Table 2　Comparison of human action recognition methods

方法	输入	准确率/%	
		UCF-101	HMDB-51
iDT[19]	RGB+OF	85.9	57.2
Two-stream[21]	RGB+OF	88.0	59.4
TSN[23]	RGB+OF	94.2	69.4
Two-stream fusion[24]	RGB+OF	92.5	65.4
STRNet+iDT[26]	RGB+OF	94.6	70.3
Hierarchical rank pooling[27]	RGB	91.4	66.9
Dynamic Image Networks+iDT[29]	RGB+OF	95.5	72.5
ActionVLAD[32]	RGB+OF	92.7	66.9
C3D[36]	RGB	82.3	51.6
Res3D[39]	RGB	85.8	54.9
LTC[40]	RGB+OF	91.7	64.8
P3D[43]	RGB+OF	88.6	—
I3D[44]	RGB+OF	97.9	80.2
ARTNet[46]	RGB	94.3	70.9
LRCN[55]	RGB+OF	82.92	—
LSTM+Two-stream[56]	RGB+OF	90.1	—
iDT+VideoLSTM[57]	RGB	91.5	63.0

从表 2 中可以看出，在国内外研究者的共同努力下，基于深度学习的方法发展得越来越迅速，识别的准确率也得到了较大的提高。目前，光流图像与 RGB 图像结合作为输入并利用大尺度数据集 Kinetics 做预训练处理的方法的识别效果最好。随着新的大尺度数据集 Kinetics 的出现、行为数据规模的扩大以及行为复杂性的提高，对人体行为识别的准确率和效率还有待进一步的提高。

5　未来研究方向

目前，基于传统手工特征的行为识别方法由于难以选取有效的特征描述符，无法应用于复杂的视频场景；而视频人体行为识别领域虽然受益于深度学习的发展而取得了巨大的进展，但仍然存在一些挑战。

1)受 GPU 和 CPU 等硬件的限制，基于深度学习的行为识别方法不能将整个视频直接输入网络模型中提取特征，只能利用连续帧间的信息冗余性从视频中提取部分帧来代表整个视频，从而提取深度特征。目前已有的研究方法大多依然使用整张图像进行特征提取，无法很好地区分前景和背景以及全局的运动信息和局部人体的运动信息，并且可能存在丢失关键动作信息的缺陷。Wang[60]通过计算相邻帧间任意两个位置的关系来直接捕捉长时依赖，而与位置距离无关，能够有效地关注视频中的运动对象，对视频非局部建模。Du 等[61]提出的交互感知时空金字塔注意力网络利用空间相邻位置的局部特征具有高相关性的特点，将多尺度特征图构建成空间金字塔，加入自注意力为每个空间位置加权，更集中关注于运动对象本身。近期，与注意力机制相结合的方法证明了注意力机制的有效性，在识别性能的提高上只是初有成效，未来还有待进一步提高。

2)目前，基于深度学习的人体行为识别方法存在的一个共同困难是需要大量的标签样本进行训练，但是在实际应用中，由于视频数据量巨大且内容多样，对视频数据进行准确、有效标注的人力和时间成本巨大，难以在产业界投入使用。佐治亚理工学院的 Ahsan 等[62]提出的 DiscrimNet，以无监督预训练的方式利用无标签数据训练一个生成对抗网络模型，再通过有标签的行为数据集微调，也取得了一定的效果。但是，由于生成对抗网络存在难以训练的问题，无监督或半监督的尝试还处于初步阶段，未来人体行为识别方法有望朝着无监督学习或半监督学习的方向发展。

3)目前基于深度学习的人体行为识别方法的网络模型计算复杂且速度较慢，在产业界难以达到实时应用的要求，因此未来期待人体行为识别方法的高效性和时效性得到更好的改进。Zolfaghari 等[63]提出的高效卷积网络，在考虑长时内容的同时快速处理每个视频，结合采样策略，利用帧间冗余性快速达到高质量的行为分类，并且整个网络模型的层数较少。目前，移动端广受大众的喜爱，但是现有的优秀方法由于网络模型较大而无法应用于移动端这样的小型设备，而目前较小的网络模型又无法达到识别性能的需求，因此能够使用较小的网络模型在移动端达到理想的效果也是未来可能的研究方向。

结束语　人体行为识别在各类生活场景中具有重要的应用价值，近年来也成为了国内外学者关注的研究热点。本文在前人工作的基础上，对人体行为识别的发展进行了综述，对基于传统手工特征的方法做出了简单的回顾，并将其与基于深度学习的方法进行了分析比较，同时介绍了目前国际上常用的人体行为识别公开数据集，最后对人体行为识别的未来发展方向做出展望，以期为初学者和未来研究方向的选择提供帮助。

参 考 文 献

[1]　HUANG K Q，CHEN X T，KANG Y F，et al. Intelligent Visual

Surveillance: A Review[J]. Journal of Computers, 2015, 38(6): 1093-1118.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet Classification with Deep Convolutional Neural Networks[C]// Proceedings of the Annual Conference on Neural Information Processing Systems. 2012: 1097-1105.

[3] AGGARWAL J K, RYOO M S. Human Activity Analysis: A Review[J]. ACM Computing Survey, 2011, 43(3): 1-43.

[4] HASSNER T. A Critical Review of Action Recognition Benchmarks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2013: 245-250.

[5] ZHU F, SHAO L, JIN X, et al. From Handcrafted to Learned Representations For Human Action Recognitions: A Survey[J]. Image and Vision Computing, 2016, 55(2): 42-52.

[6] LUO H L, WANG C J, LU F. Survey of Video Behavior Recognition[J]. Journal on Communications, 2018, 39(6): 169-180.

[7] BOBICK A F, DAVIS J W. The Recognition of Human Movement using Temporal Templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[8] FUJIYOSHI H, LIPTON A J, KANADE T. Real-time Human Motion Analysis by Image Skeletonization[J]. IEICE Transacions on Information and Systems, 2004, E87-D(1): 113-120.

[9] YANG X D, TIAN Y L. Effective 3D Action Recognition using EigenJoints[J]. Journal of Visual Communication and Image Representation, 2014, 25(1): 2-11.

[10] LAPTEV I. On Space-Time Interest Points [J]. International Journal of Computer Vision, 2005, 64(2/3): 107-123.

[11] HARRIS C J. A Combined Corner and Edge Detector[C]// Proceedings of the Alvey Vision Conferenc. 1988: 147-151.

[12] DOLLAR P, RABAUD V, COTTRELL G, et al. Behavior Recognition via Sparse Spatio-Temporal Features[C]// IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2006: 65-72.

[13] WILLEMS G, TUYTELAARS T, GOOL L. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector [C]// Proceedings of European Conference on Computer Vision. 2008: 650-663.

[14] WANG H, ULLAH M M, KLASER A, et al. Evaluation of Local Spatio-Temporal Features for Action Recognition[C]// Proceedings of the 2009 British Machine Vision Conference. 2009: 124-135.

[15] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2005: 886-893.

[16] DALAL N, TRIGGS B, SCHMID C. Human Detection using Oriented Histograms of Flow and Appearance[C]// Proceedings of the European Conference on Computer Vision. 2006: 428-441.

[17] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning Realistic Human Actions from Movies[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-8.

[18] WANG H, KLASER A, SCHMID C, et al. Action Recognition by Dense Trajectories[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2011: 3169-3176.

[19] WANG H, SCHMID C. Action Recognition with Improved Trajectories[C]// Proceedings of IEEE International Conference on Computer Vision. 2013: 3551-3558.

[20] BAY H, TUYTELAARS T, VAN GOOL L. Surf: speeded up robust features[C]// Proceedings of the European Conference on Computer Vision. Springer, 2006: 404-417.

[21] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[M]// Advances in Neural Information Processing Systems. Berlin: Springer, 2014: 568-576.

[22] WANG L M, XIONG Y J, WANG Z, et al. Towards Good Practices for Very Deep Two-Stream ConvNets[C]// Proceedings of the European Conference on Computer Vision. 2015.

[23] WANG L M, XIONG Y J, WANG Z, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition [C]// Proceedings of the European Conference on Computer Vision. 2016.

[24] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional Two-Stream Network Fusion for Video Action Recognition [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[25] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[26] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal Residual Networks for Video Action Recognition[C]// Neural Information Processing Systems. 2016.

[27] FERNANDO B, ANDERSON P, HUTTER M, et al. Discriminative Hierarchical Rank Pooling for Activity Recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[28] FERNANDO B, GOULD S. Learning End-to-End Video Classification with Rank-Pooling[C]// Proceedings of the 33rd International Conference on International Conference on Machine Learning. 2016: 1187-1196.

[29] BILEN H, FERNANDO B, GAVVES E, et al. Action Recognition with Dynamic Image Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(12): 2799-2813.

[30] GAO R H, XIONG B, GRAUMAN K. Im2flow: Motion Hallucination from Static Images for Action Recognition[J]. arXiv: 1712.04109, 2017.

[31] WANG L, GE L, LI R, et al. Three-Stream CNNs for Action Recognition[J]. Pattern Recognition Letters, 2017, 92(C): 33-40.

[32] GIRDHAR R, RAMANAN D, GUPTA A, et al. ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[33] LIN W Y, MI Y, WU J X, et al. Action Recognition with Coarse-to-Fine Deep Feature Integration and Asynchronous Fusion [C]// Thirty-Second AAAI Conference on Artificial Intelligence. North America: AAAI Publications, 2018.

[34] JI S W, XU W, YANG M, et al. 3D Convolutional Neural Networks for Human Action Recognition[C]// Proceedings of the International Conference on Machine Learning. 2010: 495-502.

[35] JI S W, XU W, YANG M, et al. 3D Convolutional Neural Net-

works for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2013，35(1)：221-231.
[36] TRAN D，BOURDEV L，FERGUS R，et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// Proceedings of IEEE International Conference on Computer Vision. 2015：4489-4497.
[37] SOOMRO K，ZAMIR A R，SHAH M. UCF101：a dataset of 101 human actions classes from videos in the wild[J]. arXiv：1212.0402，2012.
[38] KUEHNE H，JHUANG H，STIEFELHAGEN R，et al. HMDB51：a large video database for human motion recognition [C]// IEEE International Conference on Computer Vision. 2011：2556-2563.
[39] TRAN D，RAY J，SHOU Z，et al. ConvNet Architecture Search for Spatiotemporal Feature Learning[J]. arXiv：1708. 05038. 2017.
[40] VAROL G，LAPTEV I，SHMID C. Long-Term Temporal Convolutions for Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2018，40(6)：1510-1517.
[41] SUN L，JIA K ，YEUNG D Y，et al. Human Action Recognition using Factorized Spatio-Temporal Convolutional Networks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2015：4597-4605.
[42] MANSIMOV E，SRIVASTAVA N，SALAKHUTDINOV R. Initialization Strategies of Spatio-Temporal Convolutional Neural Networks[J]. arXiv：1503. 07274. 2015.
[43] QIU Z F，YAO T，MEI T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks[C]// Proceedings of IEEE International Conference on Computer Vision. 2017：5533-5541.
[44] CARREIRA J，ZISSERMAN A. Quo vadis，Action Recognition? A New Model and the Kinetics Dataset[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017：6299-6308.
[45] KAY W，CARREIRA J，SIMONYAN K，et al. The Kinetics Human Action Video Dataset[J]. arXiv：1705. 06950，2017.
[46] WANG L M，LI W，LI W，et al. Appearance-and-Relation Networks for Video Classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2018：1430-1439.
[47] HINTON G. A Practical Guide to Training Restricted Boltzmann Machines[J]. Momentum，2010，9(1)：926-947.
[48] TAYLOR G W，FERGUS R，LECUN Y，et al. Convolutional Learning of Spatio-Temporal features[C]// Proceedings of the European Conference on Computer Vision. 2010：140-153.
[49] SCHULDT C，LAPTEV I，CAPUTO B. Recognizing Human Actions：A Local SVM Approach[C]// Proceedings of the 17th International Conference on Pattern Recognition. 2004：23-26.
[50] MARSZALEK M，LAPTEV I，SCHMID C. Actions in Context [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2009：2929-2936.
[51] TRAN S N，BENETOS E，GARCEZ A. Learning Motion-Difference Features using Gaussian Restricted Boltzmann Machines for Efficient Human Action Recognition[C]// 2014 International Joint Conference on Neural Networks. 2014：2123-2129.
[52] GRAVES A，MOHAMED A，HINTON G. Speech Recognition with Deep Recurrent Neural Networks[C]// Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. 2013：6645-6649.
[53] HOCHREITER S，SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation，1997，9(8)：1735-1780.
[54] BACCOUCHE M，MAMALET F，WOLF C，et al. Sequential Deep Learning for Human Action Recognition[C]// Proceedings of IEEE International Workshop on Human Behavior Understanding. 2011：29-39.
[55] DONAHUE J，HENDRICKS L A，GUADARRAMA S，et al. Long-Term Re-current Convolutional Networks for Visual Recognition and Description [C]// The IEEE Conference on Computer Vision and Pattern Recognition. 2015：2625-2634.
[56] WU Z X，WANG X，JIANG Y G，et al. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification[C]// Proceedings of the ACM international Conference on Multimedia (ACM MM). 2015：461-470.
[57] LI Z Y，GAVRILYUK K，GAVVES E，et al. VideoLSTM Convolves，Attends and Flows for Action Recognition[J]. Computer Vision and Image Understanding，2018，166：41-50.
[58] GORELICK L，BLANK M，SHECHTMAN E，et al. Actions as Space-Time Shapes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2007，29(12)：2247-2253.
[59] RODRIGUEZ M D，AHMED J，SHAH M. Action MACH a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2008：1-8.
[60] WANG X L，GIRSHICK R，GUPTA A，et al. Non-local Neural Networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE Press，2018：7794-7803.
[61] DU Y，YUAN C F，LI B，et al. Interaction-aware Spatio-Temporal Pyramid Attention Networks for Action Classification[C]// Proceedings of the European Conference on Computer Vision. 2018：388-403.
[62] AHSAN U，SUN C，ESSA I. DiscrimNet：Semi-Supervised Action Recognition from Videos using Generative Adversarial Networks[J]. arXiv：1801. 07230，2018.
[63] ZOLFAGHARI M，SINGH K，BROX T. ECO：Efficient Convolutional Network for Online Video Understanding[C]// Proceedings of the European Conference on Computer Vision. Munich：Springer，2018：695-712.

CAI Qiang，born in 1969，Ph.D，professor，Ph.D supervisor，is a member of China Computer Federation. His main research interests include computer graphics，scientific visualization and intelligent information processing.

DENG Yi-biao，born in 1994，postgraduate. His main research interests include computer vision and human action recognition.