计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900176

基于深度学习的行为识别算法综述

赫 磊 邵展鹏 张剑华 周小龙
浙江工业大学计算机科学与技术学院 杭州 310023
(1434347689@qq.com)

摘 要 行为识别是计算机视觉领域的基本问题之一,基于深度学习的行为识别算法是当前行为识别的主流算法。在已有的研究中,传统特征提取方法一般是通过人工观察和设计,手动设计出能够表征视频动作的特征。然而,在手工特征表达的基础上构建复杂分类模型的方法已经不能适应高识别精度和应用性的要求,而深度学习的引入为行为识别带来了新的发展方向。文中主要综述了基于深度学习的行为识别算法,首先介绍了行为识别的研究背景和意义,并分别对行为识别的传统学习方法和深度学习方法进行了介绍;然后对深度学习下的算法模型结构进行分类介绍,包括Two-Stream、3D-ConvNet、融合CNN-LSTM 3种算法模型结构;最后介绍了目前常用的公开验证数据集,并主要针对基于两种数据模态的识别算法进行了横向比较,一种是基于RGB视频的UCF101和HMDB51数据集,一种是基于人体骨架序列视频的NTU RGB+D数据集。实验结果表明:深度学习方法已经取得了很大的进步,卷积神经网络的应用极大地促进了行为识别算法的发展,逐步替代了基于手工提取特征的传统方法,尤其采用了卷积神经网络算法之后在行为数据集上的准确率有了显著提高。对于RGB视频而言,Two-Stream和3DConvNet是算法模型结构的主流,对于骨架序列视频而言,Two-Stream和融合时空图模型是算法模型结构的主流。

关键词:行为识别;深度学习;卷积神经网络;循环神经网络;3D卷积
中图法分类号 TP391.4

Review of Deep Learning-based Action Recognition Algorithms

HE Lei,SHAO Zhan-peng,ZHANG Jian-hua and ZHOU Xiao-long
College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China

Abstract Action recognition is one of the fundamental problems in the field of computer vision. Currently,deep learning-based method is one of the mainstream methods for action recognition. In the existing researches,the traditional feature extraction method generally manually designs features that can represent video actions. However,this method usually requires a particular model to classify features,which cannot achieve high performance in real applications,while the introduction of deep learning brings a new development direction for action recognition. This paper briefly reviews on the action recognition methods based on deep learning. Firstly,the research background and significance of action recognition are introduced,and the traditional methods and deep learning-based methods are surveyed respectively. Then,the model architectures of three algorithms based on deep learning are classified and introduced,namely Two-Stream network,3DConvNet,CNN-LSTM network. Finally,the common used public validation datasets are introduced,and horizontal comparison is carried out on the recognition algorithms based on two data modes. Among these datasets,they can be grouped into two categories,RGB-based (e.g.,UCF101,HMDB51) and skeleton-based datasets (e.g.,NTU RGB+D). Experimental results show that the deep learning-based methods have made great advances,and the application of convolutional neural network has greatly promoted the development of action recognition algorithm. They gradually replace the traditional method based on manual features extraction. For RGB-based action recognition,Two-Stream and 3DConvNet are currently state-of-the-art methods. For skeleton-based action recognition,Two-Stream and spatiotemporal graph network achieve the best performance.

Keywords Action recognition,Deep learning,Convolutional neural network,Recurrent neural network,3D-ConvNet

1 引言

视频序列中人体动作的自动分析、检索和识别是计算机视觉中一个重要的任务,是机器视觉、模式识别、人工智能等多个学科领域的交叉研究课题,在视频监控[1-2]、人机交互[3-4]、智能机器人[5-6]、虚拟现实[7]等领域被广泛应用。人体动作的分析、检索和识别一般建立在动作特征表示的基础上[8],例如在动作识别中,在动作特征提取的基础上通过分析人体动作运动模式,建立视频序列与动作类型描述之间的映射关系,使得计算机或其他智能体可以理解视频中的动作内

基金项目:国家自然科学基金(20160283,61603341);浙江省自然科学基金(KYY-ZX-20190013,KYY-ZX-20180114)
This work was supported the National Natural Science Foundation of China (20160283,61603341) and Natural Science Foundation of Zhejiang Province,China (KYY-ZX-20190013,KYY-ZX-20180114).
通信作者:邵展鹏(zpshao@zjut.edu.cn)

容。因此,从视频序列中提取有效的动作特征以及表示特征是动作分析、检索和识别中重要的一环,直接影响到其最终结果的准确性和鲁棒性。在过去的几十年里,出现了大量的视频动作识别方法和动作识别数据集。随着神经网络在图像领域(图像分类[9]、目标检测[10]、场景分类[11])的成功应用,将神经网络用于视频动作识别的研究开始兴起,从图像到视频的成功应用说明了神经网络模型的优越性。因此,基于神经网络的行为识别已成为一个活跃的研究领域。

行人动作识别(Human Action Recognition)根据动作特征模态来分类,主要有:图像人体轮廓特征(appearance)[12]、深度图(depth map)[13]、视频人体运动光流(optical-flow)[14]以及人体骨架(body skeletons)[15]。人体图像视频不仅包含了复杂的运动背景,还有光照变化、人体外貌变化等不确定性因素,这使得基于图像视频的行为识别具有一定的局限性。相比图像视频,深度图可以很好地缓解这些不确定性因素的影响。此外,神经网络对深度图的应用也备受关注。人体行为识别的任务是在给定一段包含人体单一运动的视频片段中推断出视频中的人体动作标签(如行走、奔跑、跳跃等),其算法框架如图 1 所示。

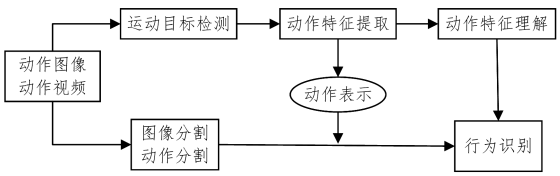


图 1 行为识别流程图

Fig. 1 Flow chart of action recognition

尽管近年来国内外人体动作行为识别的研究取得了重要进展,但人体运动的高复杂性和多变化性使得识别的精确性和高效性并没有完全满足相关行业的实用要求。实际应用中,典型的影响因素包括环境光照变化、不同视角和背景场景的动态变化等复杂外界环境干扰、视频和图像质量变化、人体身高体态的多元化等,使得人体运动的特征表示面临很多挑战,进一步影响了人体动作的建模和分析。总的来说,人体行为识别中的挑战来自以下两方面。(1)空间复杂性:不同光照、视角和背景等条件下会呈现不同的动作场景,而在不同的动作场景中相同的人体行为在姿态和特性上会产生差异。(2)时间差异性:指人体动作发生的时间点不可预测,而且动作的持续间隔也不尽相同。

深度学习下的行为识别算法按照网络架构主要划分为 3 类:双流网络、3D 卷积网络以及融合网络。本文主要分为 5 个部分:第 1 节介绍基于神经网络的行为识别的相关背景;第 2 节主要对行为识别的研究进展进行概述(按照提取特征划分:传统手工特征和深度特征);第 3 节主要对深度学习下的各种算法的模型架构进行分类介绍;第 4 节主要介绍相关数据集以及不同算法的实验对比,并通过实验得出各种方法的优缺点;最后总结全文并展望未来。

2 行为识别概述

典型的行为识别主要分 3 步进行:人体目标检测、动作表示和动作识别。人体目标检测的目的是从静态图像中分割出人体前景,从包含动作信息的视频中分割出动作序列,从而得到容量更小但包含足够运动信息的数据,并以数学符号的形式表达出人体动作。目前图像分割技术已较为成熟[16],而且人体动作行为识别越来越注重识别的实时性,简单的静态图像形式的动作识别往往出现在基础实验性质的研究中,因此动作序列的分割是未来运动目标检测的研究方向[17]。动作特征提取是为了进一步选取底层信息来实现对人体动作的表示,动作表示(也被称为特征提取)被视为视频行为识别的核心[18],动作表征选取的效果对人体动作行为识别有重要影响。有效的动作表示应该满足:(1)判别性,即来自同一类别的行为表示带有类似的信息,来自不同类别的行为表示带有区别的信息;(2)高效性,即行为表示易于计算和实现;(3)低维度,即这意味着低成本的分类和识别。最后,在动作表示提取的基础上,在空间序列和时间序列领域完成动作特征识别[19],以通过数据的分析实现动作的分类。动作特征识别可看成一个结合先验知识对数学符号进行训练和分类的过程。

行为识别按照特征提取划分的方式可以分为基于传统手工特征的行为识别和基于深度特征的行为识别。行为识别的具体划分方式如图 2 所示。传统特征提取方法一般是通过人工观察和设计,手动设计出能够表征动作特征的特征提取方法。人体行为识别特征提取方法主要分为两部分:基于人体几何或者运动信息的特征提取方法和基于时空兴趣点的特征提取方法。相比于传统方法,深度学习方法不用人工主动去提取特征,保留了视频中更多有价值的信息,效果上一般优于传统方法。深度学习方法应用在人体行为识别不仅要利用到视频的空序信息,还要用到视频的时序信息,这也是该方法研究的重点。

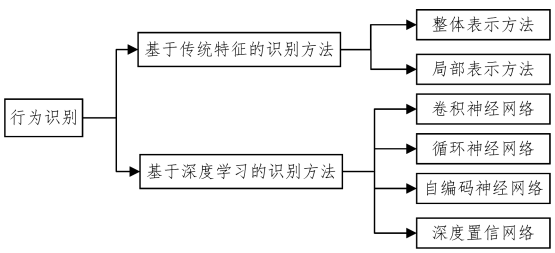


图 2 行为识别的分类

Fig. 2 Taxonomy of action recognition

2.1 基于传统特征的行为识别概述

由于缺乏对视频图像的有效表达,且硬件的配置相对落后,传统的行为识别算法主要是手工提取特征,还要建立起表示人体行为的模型,再在建立好的模型上识别人体行为,可以从表示方式上分为整体表示方法和局部表示方法。Bobick 等基于人体几何特征或运动信息的特征提取方法提出了运动能量图像(Motion Energy Images, MEI)[20]和运动历史图像(Motion History Images, MHI)[21]来解释图像序列中人的运动,其基本思想是通过一个图像对运动相关信息进行编码,通过在时空中的剪影引起的三维形状来表示动作。Dollar 等[22]的研究表明,基于整体表示方法进行编码的行为识别方法不能有效捕捉行为的视点、遮挡等变化。因此,基于人体局部特征和深度特征的研究更受青睐。对 RGB 信息提取时空兴趣点(Spatio Temporal Interest Points, STIPs)[23-24]特征已经被证明是一种有效的表示方法,它将人体动作信息以一些不关联的点的形式进行描述。目前,存在很多检测时空兴趣

点的方法，使用比较多的兴趣点检测方法包括 Harris3D 检测[25]、Hessian 检测[26]和 Cuboid 检测[27]。密集轨迹(Dense Trajectory，DT)[28]方法通过稠密光流技术在多尺度空间上跟踪特征点，提取密集轨迹点的时空兴趣点特征作为人体行为的描述子，同时采用特征词典(Bag of Features，BOF)[29]技术构建视觉词汇频率直方图表示视频行为。文献[30]对 DT 方法进行了改进，通过摄像头运动估计消除了运动产生的光流影响，然后利用 Fisher Vector[31]方法对特征进行编码，再基于编码结果训练(Support Vector Machines，SVM)[32]分类器。

基于光流的行为识别方法是深度学习领域外效果最好的方法，由 DT 算法发展而来。DT 算法的基本思路是利用光流场来获取视频序列中的一些轨迹，再沿着轨迹提取特征，最后对其特征进行编码，基于编码结果训练分类器。DT 算法的优点在于稳定性和可靠性比较高，但速度较慢。

2.2　基于深度特征的行为识别概述

在手工特征表达的基础上构建复杂模型的方法已经不能适应高精度和高速度的要求，而深度学习的引入为行为识别带来了新的发展方向。利用深度学习模型去自动提取数据中的特征，避免了人工设计特征过程中的盲目性和差异性[33]。深度学习模型通过端到端的神经网络结构进行深度特征提取和动作分类。

近年来，深度神经网络在图像处理问题上取得了巨大成果，有学者将其应用到视频分析中来解决行为识别问题[34]。在深度学习的诸多模型中，最受重视的就是被用在图像识别任务中的卷积神经网络(Convolutional Neural Networks，CNN)[35]，简称 ConvNets。循环神经网络(Recurrent Neural Networks，RNNs)[36]由于对时序建模具有优越性，已经在众多自然语言处理中(Natural Language Processing，NLP)[37-38]取得了巨大成功并被广泛应用。Wang 等提出了一种比较简单的方法[39]，通过将关节轨迹(距离)及其动态信息编码为颜色纹理图案[40]，称作关节轨迹图(Joint Trajectory Maps，JTM)，3 个 JTMS 在 3 个正交平面中生成，并提供彼此互补的信息，同时采用卷积神经网络学习区别特征进行人类行为识别，最后通过倍增分数融合[41]进一步提高了识别率。

RNN 是一个具有循环的网络，可以理解成对同一神经网络的重复操作，其允许了信息的传递性。梯度消失和梯度爆炸问题，导致标准的 RNN 对长序列进行学习时存在很多问题。为此，Hochreiter 等人提出了一个新的循环神经网络单元(Long Short-Term Memory，LSTM)[42]，旨在避免长期依赖问题。Donahue 等提出的将 CNN 与 LSTM 相结合的网络[43]，将预处理的深度图像数据先送入原先设计好的 CNN 中获取空间特征，然后将视频数据中的光流信息送入 LSTM 中获取时序特征，最终融合空序特征和时序特征并采用 Softmax 映射类别。基于人体三维骨架动作表示[44]的研究受到越来越多的关注，Shao 等提出了身体部分动作识别的层次模型[45]，按照人体的运动特性将一个人体骨架分解为多个运动刚体，提出旋转速度不变量描述子 RRV(Rotation and Relative Velocity)来表示骨架中每个刚体的旋转和速度不变量，得到动作表示。

3　深度学习下的算法模型结构

使用深度学习方法解决视频中行为识别(动作识别)的问题有两大类思路：1)以抽取并分类时空特征为目的的视频识别方法；2)以提取骨架信息进行再训练为目的的姿态估计方法。由于神经网络可以从数据中学习到特征，这种学习方式也符合人类认识世界的机理，因此，通过神经网络学习到的语义特征往往可以用于行为识别。神经网络模型按照网络结构主要划分为 3 个分支：双流(Two-Stream)方法、3D-ConvNet 方法以及融合(CNN-LSTM)方法。

3.1　Two-Stream 结构

基于 Two-Stream 模型结构的基本原理为对视频序列中每两帧计算密集光流，得到密集光流的序列(Temporal 信息)；然后对视频图像(Spatial)和密集光流(Temporal)分别训练 CNN 模型，两个分支的网络分别对动作的类别进行判断；最后直接对两个网络的训练结果进行融合，得到最终的分类结果。Two-Stream 结构的优点在于精度高，但速度慢。

Simonyan 等[46]提出的双流网络(Two-Stream Network)采用两个分支的网络架构，分别捕捉视频的空间和时间信息。空域利用 RGB 图像作为输入提取外观特征，时域利用光流信息作为输入提取时序特征，并通过多任务训练的方法对两个行为识别数据集进行分类，去除过拟合，进而获得更好的效果，如图 3 所示。这是目前的基准之一，许多网络结构也是在此基础上进行的后续探索。

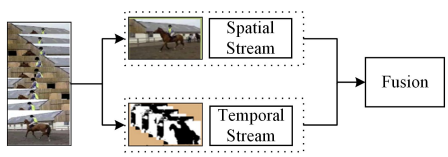


图 3　双流网络架构

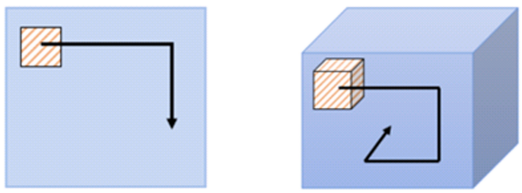
Fig. 3　Two-stream network architecture

Feichtenhofer 等[47]沿袭了双流网络的架构。为了更好地利用双流模型中的时空信息，作者在空域和时域的融合策略上进行了改进，在空域网络和时序网络的融合问题上提出了 5 种不同的融合方案，并在时域融合的问题上介绍了 3 种方法。Wang 等[48]列举了几种最新的 CNN 网络(GoogLeNet[49]，VGG-16[50])和训练策略，对比了不同 CNN 架构下双流网络的准确率。Xiong 等[51]发现前人的研究成果只能处理短期动作(short-term)，对长期动作(long-range)时间结构理解不足，且训练样本较小，于是作者使用了稀疏时间采样策略和基于视频监督的策略，将视频进行时域分割后随机抽取片段来弥补第一个不足，用交叉预练、正则化技术和数据扩张技术弥补第二个不足，并将此网络结构命名为时域分割网络(Temporal Segment Network，TSN)。由于最近残差网络(Residual Network，ResNets)[52]在深度网络训练的成功应用，Feichtenhofer 等[53]结合 ResNets 和双流模型提出了新颖的时空残差网络模型，通过空域流和时域流的残差连接分层地学习了行为的时空特征。

3.2　3D-ConvNet 结构

在行为识别中应用卷积神经网络的方法就是对视频中每一帧运用 CNN 来识别，但是这种方法并没有考虑到连续帧间的运动信息，仅能捕获到帧内的空间信息。为了有效地综合运动信息，很多文献[54-56]都用到了一种 3D 卷积的方法。通过在 CNNs 的卷积层进行 3D 卷积，捕捉在时间和空间维

度都具有区分性的特征。3D 卷积是通过堆叠多个连续的帧组成一个立方体，然后在立方体中运用 3D 卷积核。在这个结构中，卷积层中每一个特征图都会与上一层中多个邻近的连续帧相连，因此捕捉运动信息。图 4 可视化了 2D 卷积与 3D 卷积的不同。



(a)2D 卷积 (b)3D 卷积

图 4 2D 卷积与 3D 卷积的比较

Fig. 4 Comparison of 2D convolution and 3D convolution

3D-ConvNets 方法是对 2D 卷积的扩展，添加了时间维度，采用 3D 卷积核在输入视频的 3D 空间上进行采样，提取视频的时空特征，然后利用 SVM 分类器进行分类。Ji 等[57]提出了一种新的 3D-ConvNet 动作识别模型，该模型通过执行 3D 卷积从空间和时间维度中提取特征，从而捕获多个相邻帧中编码的运动信息。该模型从输入帧中生成多个信息通道，最终的特征表示组合来自所有通道的信息。Tran 等[58]提出了一种简单而有效的方法对视频数据集进行时空特征学习，对 3D-ConvNet 架构的所有网络层均采用 3×3×3 卷积核尺寸，这种网络结构被命名为 C3D。C3D 网络结构包含 8 个卷积层(3×3×3 卷积核)、5 个池化层和两个全连接层，提取的 3 维特征被证明是通用的。此外，Tran 等[59]在深度残差学习网络(ResNet18-style)中执行了 3D 卷积，所提出的架构被命名为 Res3D，其在识别精度方面要高于 C3D。为了提取视频数据的时空特征，从 2D 到 3D，利用图像识别技术处理视频识别问题是一种常见的思路，所以基本思路都是从 2D 的 CNN 结构向时域上扩展。Qiu 等[60]从另外一个角度研究了卷积核的尺寸设计，从 Inception-v3[61]中获得灵感，根据 1×3 和 3×1 的 2D 卷积核可以替代 3×3 的卷积核，并且计算量更小的思想，作者提出了所谓的伪 3D 网络，并通过将 3D 卷积拆成了 2D 卷积和 1D 卷积以及不同的串并联关系，验证了其猜测的正确性。鉴于之前 3D-CNN 最主要的缺点在于对长时时域信息没有充分挖掘，而主要问题在于网络参数多，需要大规模标注的数据集并依赖于光流，Diba 等[62]为了捕捉长时视频高层语义信息，创新性地提出了时域 3D 卷积核，并新增了时域变换层 TTL(Temporal Transition Layer)来替换池化层。整个网络被命名为 T3D(Temporal 3D-ConvNets)，且为端到端训练网络。目前的研究显示，3D-ConvNets 网络结构不及双流网络结构的识别精度高。

3D-ConvNet 结构通过 3D 卷积核去提取视频数据的时间和空间特征，这些 3D 特征提取器在空间和时间维度上操作，因此可以捕捉视频流的运动信息。3D-ConvNet 结构的最大优势在于其速度，用 Nvidia 1080 显卡可以达到 600fps 以上，所以 3D-ConvNet 的效率远远高于其他方法，这使得 3D-ConvNet 有着很好的应用前景，但 3D-ConvNet 的识别精度一般低于 Two-Stream 网络结构。

3.3 融合 CNN-LSTM 结构

融合 CNN-LSTM 网络的重点在于提取视频数据中的时空信息。很多文献中采用图卷积神经网络(Graph Convolutional Neural Networks，GCN)[63-64]提取视频中的空间特征，然后利用循环神经网络中的 LSTM 提取视频中的时序信息，融合 CNN-LSTM 结构可以理解成电路中的串联结构，这种网络结构在早期得到了广泛应用，且识别的精度较高。Karpathy 等[65]研究了时空网络中的几种融合方式，比如晚融合、早融合和慢融合，所有的融合方式如图 5 所示，这样能够获取视频中的时序信息。研究结果表明，慢融合要比早融合和晚融合的效果更好。Donahue 等[66]提出了 LRCN(Long-term Recurrent Convolutional Networks)网络，作者意识到对视频进行分析处理的关键在于对时序特征的学习和理解，故将 CNN 与 LSTM 相结合来提取视频数据中的时空信息。

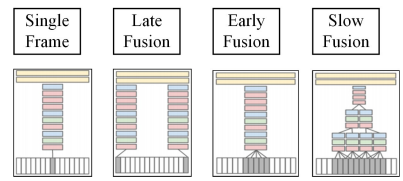


图 5 时空网络的融合方式

Fig. 5 Fusion of spatiotemporal network

Ng 等[67]认为 CNN 中的图像识别所获取到的信息是不完整的，在某些细粒度区分的场景中很容易混淆类别，为此提出了一种能够表达全局视频的 CNN 架构描述符，在实现细节上采用了时域共享参数以及光流的方法。Srivastava 等[68]认识到视频需要更高维度的特征去表达，从而需要收集更多带标签的数据并进行大量的特征提取工作，其中一个解决思路是引入非监督学习去发现、表达视频结构，可以节省给数据打标签的繁琐工作。图卷积网络模型受到越来越多的关注，Yan 等提出了时空图卷积网络模型[69]——一个基于图的动态骨骼建模方法，这是首个用以完成行为识别的基于图形的神经网络的应用。作者首先通过图卷积提取骨架的空间特征，然后通过时间卷积获取时序特征，最后融合特征，得出实验结果。Si 等[70]提出了一个新颖的网络结构 SR-TSL(Spatial Reasoning and Temporal Stack Learning)，由于现有的方法利用 LSTM 网络直接对整个骨架进行建模，利用最后一时刻的隐藏状态作为时序特征，但是对于长时序序列，最后一时刻的状态很难表示整个序列的时序特征。作者首先使用残差图神经网络对骨骼进行建模，然后将骨架序列进行等长分割送入 LSTM 中进行训练。Li 等[71]提出了端到端的卷积共现特征学习框架，使用分层方法学习共现特征，然后聚合不同级别的上下文信息，最后通过分类器进行分类识别。

不同于 Two-Stream 结构和 3D-ConvNet 结构，CNN-LSTM 结构的输入数据模态一般为骨架序列。首先通过 RGB 图像进行关节点估计或利用深度摄像机获取人体骨架序列，每一帧对应人体关节的坐标位置信息，一个时间序列由若干帧组成；然后用 CNN-LSTM 网络对构建出的骨架时序图提取高层特征；最后用 Softmax 分类器进行分类。CNN-LSTM 结构的优点在于识别精度高、算法速度较快，是目前行为识别领域最为主流的识别算法。表 1 总结了各模型结构的优缺点。

表 1　模型结构的比较

Table 1　Comparison of model structures

模型结构	介绍	优缺点
Two-Stream	对视频序列中每两帧计算密集光流，得到密集光流的序列；然后对视频图像和密集光流分别训练 CNN 模型，两个分支的网络分别对动作的类别进行判断；最后直接对两个网络的训练结果进行融合，得到最终的分类结果	精度高，速度慢
3D-ConvNet	3D-ConvNet 结构通过 3D 卷积核提取视频数据的时间和空间特征，这些 3D 特征提取器在空间和时间维度上操作，因此可以捕捉视频流的运动信息	速度快，精度低
CNN-LSTM	首先通过 RGB 图像进行关节点估计或利用深度摄像机获取人体骨架序列，每一帧对应人体关节的坐标位置信息，一个时间序列由若干帧组成；然后用 CNN-LSTM 网络对构建出的骨架时序图提取高层特征；最后用 Softmax 分类器进行分类	精度高，速度快

4　实验对比及相关数据集

4.1　常用的公开数据库

在评价不同识别方法的性能方面，现在已经有很多公开的行为数据集供研究人员使用。Hassner 等[72]综述了当前行为数据集的识别精度基准。

表 2 列出了当前公开的数据库，并提供了类别数、志愿者人数、样本数以及数据模态。

表 2　常用的公开数据集

Table 2　Common public datasets

数据库	年份	类别数	样本数量	数据模态	链接
sports-1M[73]	2014	487	120 万	RGB	http://cs.stanford.edu/people/karpathy/deepvideo
UCF101[74]	2013	101	13320	RGB	https://www.crcv.ucf.edu/data/UCF101.php
HMDB51[76]	2011	51	6766	RGB	http://serre-lab.clps.brown.edu/resource
Kinetics[77]	2017	400	300000	RGB	http://yjxiong.me/others/kinetics_action/
MSR-Action3D[78]	2010	20	567	深度图＋关节点	http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/
NTU RGB＋D[79]	2016	60	56000	RGB＋深度图＋关节点	http://rose1.ntu.edu.sg/datasets/actionrecognition.asp

UCF 包含一系列数据库，这些数据集来自 BBC/ESPN 的广播电视频道以及 YouTube 视频网站。UCF101 数据库是 UCF50[75]数据库的延伸版本，其中包含 13320 个视频样本和 101 个动作类别。UCF101 动作数据集在相机运动、物体姿态、视点变化、杂乱背景、照明条件等方面存在很大的变化，每个动作类别中的视频被分为 25 组，每组包含 4～7 个动作视频。UCF101 数据库主要包含 5 大类动作：1）人和物体交互；2）只有肢体动作；3）人与人交互；4）玩音乐器材；5）各类运动。在训练测试的过程中，建议其中的 18 组用于训练，剩余的组用于测试。

HMDB51 数据库是从各种互联网资源和数字化电影中收集的，此数据集中的人为动作主要是日常行为。该数据集中的一些关键挑战主要是摄像机视点和运动的变化、背景杂乱、志愿者位置和外观的变化。HMDB51 包含 51 个不同的动作类别，每个动作类别包含至少 101 个剪辑，总共 6766 个视频剪辑、此动作类别主要分为 5 种类型：1）一般面部动作；2）交互的面部动作；3）一般的身体动作；4）物体交互动作；5）人体交互的身体动作。对于每个动作类别，视频剪辑被分成具有 70 个剪辑的训练集和具有 30 个剪辑的测试集，并且训练和测试集中的剪辑不能来自同一个视频文件。

NTU RGB＋D 数据库是通过微软第二代 Kinect 相机进行采集的，提供了人体关节的 25 个关节节点，通过 40 个志愿者进行采集，包含 60 个动作类别，56880 个样本数据。该数据集包含了彩色图像、深度图序列、3D 骨架以及红外视频。此动作识别数据库主要分为 3 种类别：1）日常行为动作；2）医疗健康相关动作；3）双人交互动作。该数据库的评测基准被分成 Cross-Subject Evaluation 和 Cross-View Evaluation 两种，Cross-Subject Evaluation 是选取 40 个志愿者中部分人群进行训练，剩余人群用于测试；Cross-Vies Evaluation 是选取 cameras 2 和 cameras 3 用于训练，camera 1 用于测试，从目前识别精度看，一般交叉视角的精度要比交叉物体的精度高。

4.2　各方法识别性能的比较

本文主要对不同数据模态下的行为识别算法进行比较。对于原始视频 RGB 和光流（Optical Flow，OF），主要在 UCF101 和 HMDB51 数据集下进行了算法的对比；对于人体骨架 Skeleton，主要在 NTU RGB＋D 数据集下进行了算法的比较。表 3 主要描述了 RGB 和 OF 数据模态下不同识别算法的比较，表 4 主要描述了 Skeleton 数据模态下不同识别算法的比较，其中，平均精确率 *mAP* 为行为识别算法的评判标准。

表 3　RGB 和 OF 数据模态下不同识别算法的比较

Table 3　Comparison of different recognition algorithms in RGB and of data modes

Method	UCF101 *mAP*/%	HMDB51 *mAP*/%	Model Structure	Input of Network
Simonyan[46]	88.0	59.4	Two-Stream	RGB＋OF
Feichtenhofer[47]	92.5	65.4	Two-Stream	RGB＋OF
Wang[51]	94.2	69.4	Two-Stream	RGB＋OF
Pinz[53]	**94.6**	**70.3**	Two-Stream	RGB＋OF
Du[59]	85.8	54.9	3D-ConvNet	RGB
Diba[62]	93.2	63.5	3D-ConvNet	RGB
Torresani[80]	**97.3**	**78.7**	3D-ConvNet	RGB
Srivastava[68]	75.8	**44.0**	CNN-LSTM	RGB＋OF
Ng[67]	88.6	—	CNN-LSTM	RGB＋OF

表 4　Skeleton 数据模态下不同识别算法的比较

Table 4　Comparison of different recognition algorithms in skeleton data mode

Method	X-Sub *mAP*/%	X-View *mAP*/%	Structure of Network
Li[40]	76.2	82.3	Four-StreamCNN
Zhang[81]	88.2	93.8	Two-Stream RNN＋CNN
Lei[82]	**88.5**	**95.1**	Two-Stream GCN
Yan[69]	81.5	88.3	时空（GCN）
Si[70]	84.8	92.4	时空（RGNN＋LSTM）
Si[83]	**89.2**	**95.0**	时空（GCN＋LSTM）

表3比较了当前较为主流的行为识别算法的平均精确率，所用到的数据模态大多为RGB和光流OF，其中前4种的算法模型结构为Two-Stream，对UCF101数据集的平均精确率依次为88.0，92.5，94.2以及94.6，对HMDB51数据集的平均精确率依次为59.4，65.4，69.4以及70.3。中间3种的算法模型结构为3D-ConvNet，对UCF101数据集的平均精确率依次为85.8，93.2以及97.3，对HMDB51数据集的平均精确率依次为54.9，63.5以及78.7。最后两种算法模型结构为融合CNN-LSTM，对UCF101数据集的平均精确率依次为75.8和88.6。

结果表明，基于深度学习的行为识别算法的识别精度得到了明显的提升。对于算法的模型结构而言，由于Two-Stream结构和3D-ConvNet结构可以对视频序列获取时序信息，目前Two-Stream和3D-ConvNet两种模型结构的识别精度较高。在Two-Stream结构中，Simonyan使用了RGB和OF分别作为输入，得到了不错的效果。Feichtenhofer在此基础上改变了不同的融合方式，对UCF101和HMDB51的识别精度分别提高了4.5%和6%。Wang使用了稀疏时间采样策略和基于视频监督的策略，对UCF101和HMDB51的识别精度分别提高了1.7%和4%。Pinz使用了时空残差网络ResNet对光流进行时域卷积，时域网络和空域网络由残差连接进行参数传递，提高了识别精度。在3D-ConvNet结构中，Du是C3D算法的创始人，作者对不同输入帧数、网络层数、分辨率、卷积核感受野大小做了详细的对比。Diba发现3D卷积的缺点在于对长时时域信息没有充分挖掘，于是作者创新地提出了新的时域3D卷积核，对数据集的识别精度有了显著提升。Torresani提出接近输入端的浅层次用3D卷积进行训练，深层次用2D卷积进行训练，且都使用残差卷积网络来进行学习，对UCF101和HMDB51的识别率达到了最高。综上所述，对于数据模态为RGB和光流来说，Two-Stream结构和3D-ConvNet结构是算法模型结构中的主流。结果表明，目前基于人体骨架的行为识别算法的模型结构大多为分流模型结构和时空融合模型结构，对三维人体骨架很少有使用3D-ConvNet模型结构的。Li将三维骨架编码成了4个颜色纹理图案，然后进行分别卷积，最后融合。Zhang将人体骨架进行分流处理，分别将人体骨架序列送入CNN和RNN网络中，最后融合分类。Lei在时空图卷积的基础上利用每一帧的骨架轮廓和每一帧的关节点形成Two-Stream分流，然后分别使用图卷积，最后融合，是目前Two-Stream结构中识别精度比较高的模型，在X-sub和X-view的精度分别为88.5%和95.1%。

由于三维人体骨架可以克服视频图像中复杂的运动背景、光照变化等因素影响，目前对人体骨架的研究受到越来越多的关注，对人体骨架的算法模型结构主要为时空融合模型结构。Yan使用了时空图卷积对人体骨架进行提取特征，开启了图卷积在骨架的应用。Si改进了时空图卷积模型，首先使用残差图神经网络或者图卷积神经网络对人体骨架获取空间特征，然后利用LSTM获取时序信息，在NTU RGB+D数据集上达到了很高的识别精度。

结束语 本文首先对行为识别的研究背景和意义进行概述，然后对传统的行为识别方法和基于深度学习的人体行为识别方法进行了分析总结。基于深度学习的行为识别方法不需要像传统方法那样对特征的提取进行人工设计，可以在视频数据上进行训练和学习，得到最有效的特征表示。其次，对深度学习下的算法模型结构进行分类介绍。最后，介绍了行为识别相关的常用数据集以及各种识别算法在数据集下的性能比较。目前行为识别存在的问题、难点以及解决问题的思想如下。

1)视角无关的行为识别研究方法。在实际的应用中，视频都是以任意视点采集视频的，所以在实际的应用中对视点无关的要求非常高。现在大部分数据库都是在单一视角或者按照人体正面、侧面进行采集的，因此，未来需要研究一种识别方法可以应对复杂的视角变化，另外还需要提供多视角的数据库为各种识别算法提供一个基准。最后，如何克服视角变化的影响也是行为识别未来研究的方向。

2)弱监督方法以及以后的无监督方法。基于监督学习的行为识别严重依赖于视频动作标签的标注，而对视频图像进行标注会耗费大量的资源。随着深度学习的不断发展，视频图像标注的成本变得越来越高，如何利用低成本的图像标注取得良好的识别结果成为了当下研究的热点。因此，研究者开始研究基于弱监督及无监督的行为识别算法，由无需人工标签或者标签少量，具有很大的应用价值，是未来非常有前景的研究方向之一。

3)行为识别严重依赖物体和场景。行为识别大多采用早期的行为数据集，大多数行为为跑步、骑马、遛狗滑雪类，这些数据对算法带来了一些导向。比如，跑步和骑马的区别是什么？就是一匹马，那直接用马的检测器就可区分。对于遛狗和滑雪，看背景是什么场景就可区分。因此，算法越来越偏向用物体和场景来识别，始终没有切入到“动作”本身，这个问题不局限于图片，视频中也存在。

参 考 文 献

[1] WANG X. Intelligent multi-camera video surveillance: A review [J]. Pattern Recognition Letters, 2013, 34(1): 3-19.

[2] TURAGA P, CHELLAPPA R, SUBRAHMANIAN V S, et al. Machine Recognition of Human Activities: A Survey[J]. IEEE Trans. Circuits Syst. Video Technol., 2008, 18(11): 1473-1488.

[3] ELLIS C, MASOOD S Z, TAPPEN M F, et al. Exploring the trade-off between accuracy and observational latency in action recognition[J]. Int. J. Comput. Vis., 2013, 101(3): 420-436.

[4] ZHANG W, SMITH M L, SMITH L N, et al. Gender and gaze gesture recognition for human-computer interaction[C]// Computer Vision and Image Understanding. 2016: 32-50.

[5] TAKANO W, NAKAMURA Y. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions[J]. Int. J. Rob. Res., 2015, 34(10): 1314-1328.

[6] CALO R. Robotics and the Lessons of Cyberlaw[J]. California Law Review, 2014, 103(3).

[7] CAMPORESI C, KALLMANN M, HAN J J, et al. VR solutions for improving physical therapy[C]// IEEE Virtual Reality Conference. 2013: 77-78.

[8] CHAO M W, LIN C H, ASSA J, et al. Human motion retrieval from hand-drawn sketch[J]. IEEE Trans. Vis. Comput. Graph., 2012, 18(5): 729-740.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet

classification with deep convolutional neural networks[C]// Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS). 2012:1097-1105.

[10] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014:580-587.

[11] FARABET C, COUPRIE C, LECUN Y. Learning hierarchical features for scene labeling[J], IEEE Trans. Pattern Anal. Mach. Intell., 2013, 35(8):1915-1929.

[12] SU S, LIU Z, XU S, et al. Sparse auto-encoder based feature learning for human body detection indepth image[J]. Signal Processing, 2015, 112(1):43-52.

[13] VIEIRA A W, NASCIMENTO E R, OLIVEIRA G L, et al. STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences[C]// Iberoamerican Congress on Pattern Recognition. 2012:252-259.

[14] SEVILLALARA L, LIAO Y, GUNEY F, et al. On the Integration of Optical Flow and Action Recognition[C]// German Conference on Pattern Recognition. 2018:281-297.

[15] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. 2016:770-778.

[16] TSENG H C, SHYU J J, CHANG J Y, et al. Exploiting Automatic Image Segmentation to Human Detection and Depth Estimation[C]// /Proc of the IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Processing. Paris, France, 2011:19-25.

[17] KIM W H, JEONG T I, KIM J N. Video Segmentation Algorithm Using Threshold and Weighting Based on Moving Sliding Window[C]// Proc of the 11th International Conference on Advanced Communication Technology. Pyeongchang County, Repulic of Korea, 2009:1781-1784

[18] SALMANE H, RUICHEK Y, KHOUDOUR L. Object Tracking Using Harris Corner Points Based Optical Flow Propagation and Kalman filter[C]// Proc of the 14th International IEEE Conference on Intelligent Transportation Systems. Washington, USA, 2011:67-73.

[19] YANG J, XU Y S, CHEN C S. Hidden Markov Model Approach to Skill Learning and Its Application to Telerobotics[J]. IEEE Trans on Robotics and Automation, 1994, 10(5):621-631.

[20] BOBICK A, DAVIS J. An appearance-based representation of action[C]// Proceedings of the 13th International Conference on Pattern Recognition. Vienna: IEEE, 1996:307-312.

[21] WEINLAND D, RONFARD R, BOYER E. Free viewpoint action recognition using motion history volumes[J]. Computer Vision and Image Understanding, 2006, 104(2/3):249-257.

[22] DOLLAR P, RABAUD V, COTTRELL G W, et al. Behavior recognition via sparse spatio-temporal features[C]// International Conference on Computer Communications and Networks. 2005:65-72.

[23] LAPTEV I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64 (2/3):10-123.

[24] WONG S, CIPOLLA R. Extracting Spatiotemporal Interest Points using Global Information[C]// International Conference on Computer Vision. 2007:1-8.

[25] WANG H, ULLAH M M, KLASER A, et al. Evaluation of local spatio-temporal features for action recognition[C]// British Machine Vision Conference. 2009:1-11.

[26] WANG H, KLASER A, SCHMID C, et al. Densetrajectories and motion boundary descriptors foraction recognition[J]. International Journal of Computer Vision, 2013, 103(1):60-79.

[27] DOLLAR P, RABAUD V, COTTRELL G W, et al. Behavior recognition via sparse spatio-temporal features[C]// International Conference on Computer Communications and Networks. 2005:65-72.

[28] WANG H, KLASER 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] NGUYEN T PMANZANERA A. Action recognition using bag of features extracted from a beam of trajectories[C]// 2013 IEEE International Conference on Image Processing. Melbourne, VIC, 2013:4354-4357.

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

[31] SHI C, WANG Y, JIA F, et al. Fisher vector for scene character recognition: A comprehensive evaluation[J]. Pattern Recognition, 2017, 2017(72):1-14.

[32] DANAFAR S, GHEISSARI N. Action recognition for surveillance applications using optic flow and SVM[C]// Asian Conference on Computer Vision. 2007:457-466.

[33] WANG Y, XU W. Leveraging deep learning with LDA - based text analytics to detect automobile insurance fraud[J]. Decision Support Systems, 2018, 105:87-95.

[34] IJJINA E P, MOHAN C K. Hybrid deep neural network model for human action recognition[J]. Applied Soft Computing, 2016, 46:936-952.

[35] KRIZHEVSKY A, SUTSKEVER I, HINTON G E, et al. ImageNet Classification with Deep Convolutional Neural Networks [J]. Neural Information Processing Systems, 2012, 141(5): 1097-1105.

[36] GREFF K, SRIVASTAVA R K, KOUTNIK J, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks, 2017, 28(10):2222-2232.

[37] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural Language Processing (Almost) from Scratch[J]. arXiv:1103.0398.

[38] TARWANI K M, EDEM S. Survey on Recurrent Neural Network in Natural Language Processing[J]. International Journal of Engineering Trends and Technology, 2017, 48(6):301-304.

[39] WANG P, LI Z, HOU Y, et al. Action Recognition Based on Joint Trajectory Maps Using Convolutional Neural Networks [C]// ACM Multimedia. 2016:102-106.

[40] LI C, HOU Y, WANG P, et al. Joint Distance Maps Based Action Recognition With Convolutional Neural Networks[J]. IEEE Signal Processing Letters, 2017, 24(5):624-628.

[41] WANG X, GAO L, WANG P, et al. Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length[J]. IEEE Transactions on Multimedia, 2018, 20(3): 634-644.

[42] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[43] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Computer Vision and Pattern Recognition, 2015. 2625-2634.

[44] KE Q, BENNAMOUN M, AN S, et al. A New Representation of Skeleton Sequences for 3D Action Recognition[C]// Computer Vision and Pattern Recognition. 2017: 4570-4579.

[45] SHAO Z, LI Y, GUO Y, et al. A Hierarchical Model for Action Recognition Based on Body Parts[C]// 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, QLD, 2018: 1978-1985.

[46] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[J]. arXiv: 1406. 2199.

[47] FEICHTENHOFER C, PINZ A, ZISSERMAN A, et al. Convolutional Two-Stream Network Fusion for Video Action Recognition[C]// Computer Vision and Pattern Recognition. 2016: 1933-1941.

[48] WANG L, XIONG Y, WANG Z, et al. Towards Good Practices for Very Deep Two-Stream ConvNets[J]. arXiv: 1507. 02159, 2015.

[49] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Computer Vision and Pattern Recognition. 2015: 1-9.

[50] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]// International Conference on Learning Representations. 2015.

[51] WANG L, XIONG Y, WANG Z, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition [C]// European Conference on Computer Vision. 2016: 20-36.

[52] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. 2016: 770-778.

[53] FEICHTENHOFER C, PINZ A, WILDES R P, et al. Spatiotemporal Residual Networks for Video Action Recognition[C]// Neural Information Processing Systems. 2016: 3468-3476.

[54] YANG M, JI S, XU W, et al. Detecting human actions in surveillance videos[C]// Proceedings of the TREC Video Retrieval Evaluation Workshop. 2009.

[55] BACCOUCHE M, MAMALET F, WOLF C, et al. Sequential deep learning for human action recognition[C]// Human Behavior Unterstanding. 2011: 29-39.

[56] 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.

[57] 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.

[58] TRAN D, BOURDEV L, FERGUS R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// International Conference on Computer Vision. 2015: 4489-4497.

[59] TRAN D, RAY J, SHOU Z, et al. ConvNet Architecture Search for Spatiotemporal Feature Learning. [J]. arXiv: 1708. 05038, 2017.

[60] QIU Z, YAO T, MEI T, et al. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks[C]// International Conference on Computer Vision. 2017: 5534-5542.

[61] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[C]// Computer Vision and Pattern Recognition. 2016: 2818-2826.

[62] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification[J]. arXiv: 1711. 08200, 2017.

[63] KIPF T, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]// International Conference on Learning Representations. 2017.

[64] SHI L, ZHANG Y, CHENG J, et al. Non-Local Graph Convolutional Networks for Skeleton-Based Action Recognition [J]. arXiv: 1805. 07694v2.

[65] KARPATHY A, TODERICI G, SHETTY S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// Computer Vision and Pattern Recognition. 2014: 1725-1732.

[66] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Computer Vision and Pattern Recognition. 2015: 2625-2634.

[67] NG J Y, HAUSKNECHT M J, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification[C]// Computer Vision and Pattern Recognition. 2015: 4694-4702.

[68] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R, et al. Unsupervised Learning of Video Representations using LSTMs[C]// International Conference on Machine Learning. 2015: 843-852.

[69] YAN S, XIONG Y, LIN D, et al. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition [C]// National Conference on Artificial Intelligence. 2018: 7444-7452.

[70] SI C, JING Y, WANG W, et al. Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning[C]// European Conference on Computer Vision. 2018: 106-121.

[71] LI C, ZHONG Q, XIE D, et al. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation[C]// International Joint Conference on Artificial Intelligence. 2018: 786-792.

[72] HASSNER T. A critical review of action recognition benchmarks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2013: 245-250.

[73] KARPATHY A, TODERICI G, SHETTY S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// Computer Vision and Pattern Recognition. 2014: 1725-1732.

[74] SOOMRO K, ZAMIR A R, SHAH M, et al. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. arXiv: 1212. 0402, 2012.

[75] REDDY K K, SHAH M. Recognizing 50 human action categories of web videos[J]. Machine Vision Applications, 2013, 24(5): 971-981.

[76] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A large video database for human motion recognition[C]// International Conference on Computer Vision. 2011: 2556-2563.

[77] ZISSERMAN A, CARREIRA J, SIMONYAN K, et al. The Kinetics Human Action Video Dataset[J]. arXiv: 1705. 06950, 2017.

[78] LI W，ZHANG Z，LIU Z，et al. Action recognition based on a bag of 3D points[C]// Computer Vision and Pattern Recognition. 2010：9-14.

[79] SHAHROUDY A，LIU J，NG T，et al. NTU RGB+D：A Large Scale Dataset for 3D Human Activity Analysis[C]// Computer Vision and Pattern Recognition. 2016：1010-1019.

[80] TRAN D，WANG H，TORRESANI L，et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[C]// Computer Vision and Pattern Recognition. 2018：6450-6459.

[81] ZHANG P，LAN C，XING J，et al. View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019：1-1.

[82] SHI L，ZHANG Y，CHENG J，et al. Adaptive spectral graph convolutional networks for skeleton-based action recognition [J]. arXiv：1805. 07694，2018.

[83] SI C，CHEN W，WANG W，et al. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition[J]. arXiv：1902. 09130，2019.

HE Lei，born in 1994，master. His main research interests include image processing and action recognition.

SHAO Zhan-peng，Ph.D，is a member of China Computer Federation. His research interests include action recognition and pose estimation.

（上接第 134 页）

2700，4100 个情节左右收敛。$\beta=0.2$ 的算法在训练初期学习速度较快，获得的平均奖赏波动较大；其他 β 值的算法在训练初期学习速度慢，获得的平奖奖赏较小。这是因为在训练过程中，不同的 β 取值会改变线性函数值函数的大小，从而影响损失函数，但是并不影响算法的收敛性。因此，DQN-LF 算法能在保证收敛的情况下，有较快的学习、收敛速度。

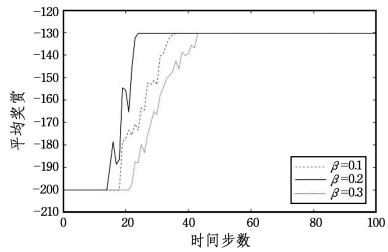


图 8　不同 β 取值下 DQN-LF 算法的比较

Fig. 8　Comparison of DQN-LF algorithms with different β values

结束语　本文主要针对 DQN 神经网络在训练前期速度慢的问题，提出一种新的基于函数逼近协同更新的 DQN 算法。该算法引入线性函数，并利用训练初期线性函数的值函数代替 DQN 网络中的行为网络的值函数作差作为损失函数，利用线性函数收敛速度快的特性，提高了训练前期参数优化的速度。实验基于 OpenAI Gym 平台，结果表明，基于函数逼近的 DQN 协调更新算法加快了训练初期的学习速度和收敛速度。

本文主要利用 OpenAI Gym 实验平台对算法做了相关分析，从结果可以看出，算法在训练初期具有较快的学习速度和收敛速度。CartPole 和 Mountain Car 是一种小规模的连续状态空间问题，接下来考虑将算法应用于具有较大规模的连续空间实际问题，进一步检验算法的性能。

参 考 文 献

[1] SUTTON R S. Learning to Predict by the Methods of Temporal Differences[J]. Machine Learning，1988，3(1)：9-44.

[2] ZHANG R B，GU G C，LIU Z D，et al. Reinforcement Learning Theory，Algorithms and Its Application[J]. Control Theory and Application，2000，17(5)：637-642.

[3] SUTTON R S，BARTO A G. Reinforcement learning：An introduction[M]. Massachusetts：MIT press，2018.

[4] WATKINS C J C H，DAYAN P. Q-learning[J]. Machine Learning，1992，8(3/4)：279-292.

[5] HINTON G E，SALAKHUTDINOV R R. Reducing the dimensional-ity of data with neural networks[J]. Science，2006，313(5786)：504-507.

[6] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level con-trol through deep reinforcement learning[J]. Nature，2015，518(7540)：529.

[7] OSBAND I，BLUNDELL C，PRITZEL A，et al. Deep exploration via bootstrapped DQN[C]// Advances in Neural Information Processing Systems. Barcelona，Spain，2016：4026-4034.

[8] ANSCHEL O，BARAM N，SHIMKIN N. Averaged-DQN：Variance Reduction and Stabilization for Deep Reinforcement Learn-ing[C]// Advances in International Conference on Machine Learning. New York，USA，2016.

[9] WANG Z，SCHAUL T，HESSEL M，et al. Dueling network architectures for deep reinforcement learning[J]. arXiv：1511. 06581，2015.

[10] LEVINE S，FINN C，DARRELL T，et al. End-to-End Training of Deep Visuomotor Policies[J]. Journal of Machine Learning Research，2015，17(39)：1-40.

[11] LEVINE S，PASTOR P，KRIZHEVSKY A，et al. Learning hand-eye coordination for robotic grasping with large-scale data collection[C]// Proceedings of International Symposium on Experimental Robotics. Berlin：Springer，2016：173-184.

LIU Qing-song，master candidate. His main research interests include reinforcement learning and building energy efficiency.

CHEN Jian-ping，doctor，professor. His research interests include big data and analytics，building energy efficiency，and intelligent information.