DOI:10.16652/j.issn.1004-373x.2020.18.039

基于3D CNN的人体动作识别研究

朱云鹏，黄　希，黄嘉兴
（南通大学 机械工程学院，江苏 南通　226019）

摘　要：在现实的生活视频中，检测人体动作以及分类时，常常会出现视频背景复杂、模糊，以及因人多导致多种动作行为同时出现的问题，而致使检测和判别某种行为结果出现偏差。因此文中针对2D CNN对单个帧进行提取特征却没有包含实际视频中连续多帧之间编码的运动信息，提出一种基于三维卷积神经网络识别方法，旨在更好地捕获视频连续帧中隐藏的时间和空间信息。实验结果表明，与现有的几类方法相比，所提方法识别率得到较为明显的提升，验证了该方法的有效性和鲁棒性。

关键词：人体动作识别；三维卷积神经网络；特征提取；模型训练；深度学习；实验对比
中图分类号：TN911.23-34；TP301　　**文献标识码：**A　　**文章编号：**1004-373X(2020)18-0150-03

Human action recognition based on 3D CNN

ZHU Yunpeng, HUANG Xi, HUANG Jiaxing
(School of Mechanical Engineering, Nantong University, Nantong 226019, China)

Abstract: In real-life video detection and classification, the video background is complex and fuzzy, as well as many people lead to a variety of action behavior problems at the same time, which causes the deviation of detection and discrimination of a certain behavior results. In allusion to the problem that feature extraction is conducted by 2D CNN from a single frame, but the motion information encoded between consecutive frames is not included, a neural network recognition method based on 3D convolution is proposed to better capture the hidden time and space information in consecutive frames of video. The experimental results show that, in comparison with the existing methods, the recognition rate of this method is significantly improved, and the effectiveness and robustness of the proposed method are verified.

Keywords: human action recognition; 3D convolution neural network; feature extraction; model training; deep learning; experimental comparison

0　引　言

人体动作识别是计算机视觉研究中的一个分支，被广泛地应用于人机互动、交互式娱乐等多个领域[1]。随着视频采集设备和宽带光纤整体科技水平的跳跃式发展，如今，“视频”已经成为信息的主要载体，特别是近些年来，4G的普及以及5G的问世，各色各类的长、短视频数量以几何速度爆炸式增加，面对如此庞大的视频数据，相关人员迫切需要稳定高效的视频信息自动处理系统。在此供求基础上，人体动作识别技术近些年来一直是计算机领域内一个充满机遇和挑战的课题。

最常见的动作识别应用是分类识别：给定一个视频，系统将其准确分类为几个已知的动作类别。综合性的动作识别是视频中不仅包含的多个动作类别，还存在复杂的背景干扰。动作识别的最终目标是分析人在视频中场景的位置、状态和行为。人体动作识别应用于各行各业，主要集中在智能视频监控、病人监护系统、人机交互、虚拟现实、智能家居、智能安全、运动员辅助培训、基于情报的视频检索和智能图像压缩等[2]。随着传感器技术的不断发展，人类行为识别研究受益于不同模态传感器，如RGB摄像机、深度摄像机、加速度计和陀螺仪[3]。

图像和视频的识别与描述是计算机视觉领域的一个基本挑战[4]。而与图像分类相比，视频动作分类在运动和视角上存在着附加的难题[5]。视觉人体运动分析和识别的方法体系有很多种，如：Forsyth等人侧重于将动作从视频序列中人的姿态和运动信息恢复过来，这属于一个回归问题，而人体行为识别是一个分类问题。这2个

收稿日期：2020-01-08　　**修回日期：**2020-03-17
基金项目：国家自然科学基金青年基金项目(51405246)；南通市科技局项目(CP12014001；MS12017017-7)

问题有很多类似点,比如其特征的提取和描述很多方面是通用的。如果将人体运动识别的研究方向分为3个层次:移动识别、动作识别和行为识别,目前关于行为识别基本上还停留在第二个阶段,即对生活中的一些简单行为进行判断和分类。与传统的模式识别方法相比,基于深度学习的人体运动识别技术近年来发展迅速,它的研究结合自动训练,提取特征和分类,同时放宽了有关参数的数量,并且利用深度学习将人体动作识别的研究投入到新的应用当中。

深度学习允许由多个处理层组成的计算模型来自动学习多维的抽象数据类型[6]。它的主要优势之一是其执行端到端优化的能力[7]。目前,使用深度学习执行诸如人体运动识别、人体跟踪和图像高级处理之类的任务均得到了令人满意的结果,如麻省理工学院媒体实验室在将智能室以及在自然场景中的人体动作识别作为新的研究课题中取得了一些进展;CMU机器人研究所还开展了人体检测与跟踪、步态识别和行为识别等项目;同时,马里兰大学自动化研究控制中心对人体运动建模,对3D人体运动捕捉和异常事件检测也进行了深入的研究。现实的生活视频中检测人体动作及分类时,会出现视频背景复杂、模糊,以及因人多导致多种动作行为同时出现的问题,致使检测和判别某种行为结果出现偏差。 本文针对2D CNN对单个帧进行提取特征却没有包含实际视频中连续多帧之间编码的运动信息,提出一种基于改进三维卷积神经网络识别方法,旨在更好地捕获视频连续帧中隐藏的时间和空间信息,并且在多个动作识别视频数据集实验中得到了较高的准确率。

1　改进三维卷积神经网络模型

1.1　3D CNN网络结构组成

受视觉神经感受野的启发,卷积神经网络的神经元之间通过稀疏链接的方式进行连接,具有较多的隐含层,每一隐含层有多个数据矩阵平面,每个数据矩阵平面的神经元共享权值参数矩阵[8]。如图1所示,在二维卷积神经网络中,卷积应用于二维特征图,并且仅根据空间维度计算特征。

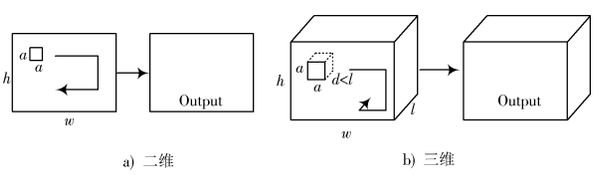

图1　二维和三维卷积过程对比

当使用视频数据分析问题时,需要在多个连续帧中捕获编码的运动信息。 为此,提出三维卷积神经网络用于计算空间和时间维度特征。三维卷积是通过堆叠多个连续帧,然后在立方体中应用三维卷积内核来形成立方体。利用这种结构,卷积层中的特征映射连接到上层中的多个相邻帧,从而捕获运动信息。深度神经网络的优势主要在于学习训练数据的分布,并且可以在测试集上获得良好的泛化效果。 然而如果每个批次输入的数据都具有不同的分布,则会给神经网络的训练带来困难,所以规范化每层神经网络的输出显然是不合理的。为了把每层神经网络任意神经元输入值的分布拉回到均值为0,方差为1的标准正态分布,本文神经层中引入批量规范化(Batchnorm)。假设神经层输入数据是$\beta=x_1,x_2,\cdots,x_m$,共m个数据,输出是$y_i=\mathrm{BN}(x)$,则批量规范化步骤如下:

1)求此次批量书的数学期望:$\mu_\beta=\sum_{i=1}^{m}x_i$;

2)求此次batch的方差,$\sigma_\beta^2=\frac{1}{m}\sum_{i=1}m(x_i-\mu_\beta)^2$;

3)接下来对x做归一化,得到$x_{\bar{i}}$;

4)引入缩放和平移变量γ和β,计算归一化后的值,$y_i=\gamma x_{\bar{i}}+\beta$

在sigmoid, tanh, softplus和ReLu中,选择ReLu作为网络的非线性激活函数,ReLu的gradient大多数情况下是常数,有助于解决深层网络的收敛问题。ReLu的另一个优势是在生物维度上的合理性,它是单边的,相比sigmoid和tanh,更符合生物神经元的特征。为了确保特征的位置和旋转不变性并减少过拟合问题,在网络中添加最大池化层,从过滤器中提取一些特征值,并且仅获取最大池化层作为保留值,丢弃所有其他功能值。在视频领域,如果在足够大的数据集上训练动作分类网络,在应用于不同的时间任务或数据集时,是否会提供相应的性能提升,这是一个悬而未决的问题[9],本文在网络的训练阶段加入Dropout技术来随机地选择部分神经元并将其输入设置为0,从而随机变化地网络的链接结构,提高网络的泛化能力,使得网络具有更好的适应性[10]。

1.2　方法实施过程

1.2.1　网络结构

从图2可以看出,该特征通过3次三维卷积和3次下采样组合,最后通过完全连接层获得最终输出。

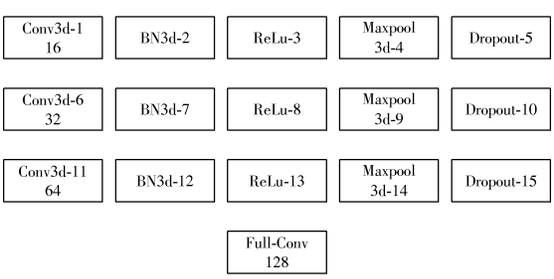

图2　网络实现过程

1.2.2 图像大小的变化

卷积过程中的尺寸变化如图3所示。

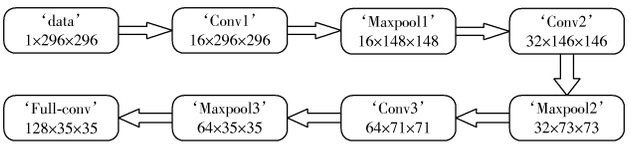

图3 卷积过程中的尺寸变化

2 实验方法

2.1 实验环境

编程环境使用Python 3.6,而Numpy,Tensorflow和其他一些模块也会被用到。选择UT交互数据集作为实验研究数据集。UT交互数据集包含6类真实的人人交互行为,包括握手、指向、拥抱、推、踢和击打。每个视频每次互动至少包含1次执行或2~3次执行。在第1~第4组中,场景中只出现2个相互作用的人。在第5~8组中,场景中存在执行其他动作的干扰人员。所有视频中出现了超过15种不同服装的参与者。本文选取数据集15人中的8人作为训练样本,7人作为测试样本。

2.2 实验过程

2.2.1 特征提取

对于每个实验视频,将其分成多组15个连续帧的块,然后在这些块上训练模型而不是在单独每一帧上训练。在卷积层中,使用3D卷积滤波器来训练模型以检测并学习时间运动信息。特征提取如图4所示。

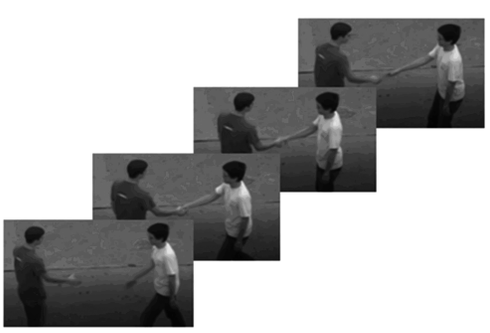

图4 特征提取

2.2.2 参数影响

1) Learning rate

学习率是深度学习中一个举足轻重的超级参数。能否选择搭配网络结构的最优学习率决定了模型塑造的质量。文中,学习率调整到0.01的获得最高准确率结果。

2) Dropout

Dropout是指在深度学习网络的训练过程中,对于神经网络单元,按照一定的概率将其暂时从网络中丢弃,是一种很有效的正则化手段[11],对于随机梯度下降来说,由于是随机丢弃,故每一个Mini-batch都在训练不同的网络,每次丢失时,都相当于从原始网络中找到更薄的网络。

2.2.3 实验结果比较

不同方法实验结果比较如表1所示。

表1 不同方法实验结果比较 %

Methods	识别率	握手	拥抱	踢人	指认	拳打	推人
Bag-of-words	68.33	50	70	80	95	50	70
Ryoo & Aggar	70.8	75	87.5	75	62.5	50	75
Yu et al	83.33	100	65	75	100	85	75
Ryoo	85	—	—	—	—	—	—
Yu kong et al	88.33	80	80	100	90	90	90
Our method	90.83	87	90	99	99	80	90

UT数据集是人与人交互式类的行为数据集,即便该数据集的动作分类单一且动作本身不具备复杂性,但由于人与人之间交互时的遮挡和不确定性,导致识别难度提高,相似动作容易混淆,分类算法准确率浮动较大。例如表1所示:本文算法在此数据集上,“拳打”动作准确率最低,只有80%,原因在于“拳打”和“推人”动作近似,算法易发生误判;除“握手”和“拳打”动作之外,其余动作识别准确性均在90%以上。可见,本文算法在UT数据集上识别率得到了一定程度上的提高。

3 结 论

动作识别系统的性能在很大程度上取决于它是否能够高效提取和利用相关信息[12]。而动态图像是紧凑的,在将视频转换成动态图像期间,时间信息在某种程度上不可避免地丢失[13]。本文通过基于改进三维卷积神经网络学习方法和其他实验方法在UT数据集得到的相比较,更大程度上利用测试视频包含的空间和时间信息,并且准确率得到了一定程度的提高,证明了该方法在短视频交互动作识别中的可行性。

注:本文通讯作者为黄希。

参 考 文 献

[1] 张孙培,孙怀江.关节信息和极限学习机的人体动作识别[J].现代电子技术,2015,38(10):55-60.

[2] 刘文婷.一种室内人体行为识别方法:CN104866860A[P].2015-08-26.

[3] CHEN Chen, ROOZBEH Jafari, NASSER Kehtarnavaz. UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor [C]// 2015 IEEE International Conference on Image Processing. Quebec City: IEEE, 2015: 168-172.

(下转第156页)

感，可以明确找出重构图像轮廓线存在的毛刺。经过处理后，即可得到完整的光滑轮廓，但在取出毛刺时，会受到断裂点周围噪声的影响。在后续研究中，要在细化算法的基础上，引入去噪增强算法，减少周围噪声的影响。在今后的研究中要从以下几点改进：

1）进一步提高细化算法的准确性；

2）寻找普适性更强的有效算法；

3）通过实验验证所提的三维动态动画人物重构中的图像毛刺消除技术设计的可行性。

参 考 文 献

[1] 路永婕，李振宇，怀文青，等.基于分形理论的三维路面谱重构及在多体动力学软件中的应用[J].图学学报，2019，40(2)：328-334.

[2] 陈卓，胡摇，蒋晓黎，等.微结构形貌的光场显微三维重构分辨率增强技术[J].光学技术，2018，44(4)：385-390.

[3] 杨小来，廖巨华.一种解决高速DAC转换毛刺的同步方法[J].科技通报，2018，34(10)：168-171.

[4] 梁建平.三维激光3D打印技术在数字化重现中的应用[J].激光杂志，2018，39(6)：125-129.

[5] 叶凤华，叶欢.基于FLASH和3D动画渲染技术的育苗机器人设计[J].农机化研究，2018，40(3)：189-192.

[6] 李忠虎，张琳，闫俊红.管道腐蚀视觉测量图像边缘检测算法研究[J].电子测量与仪器学报，2017，31(11)：1788-1795.

[7] 郑冰，赵阳，葛东林.农机快速导航系统设计：基于图像边缘检测和3D深度视频帧内编码[J].农机化研究，2018，40(5)：181-184.

[8] 马宏伟，杨文娟，张旭辉.带式输送机托辊红外图像分割与定位算法[J].西安科技大学学报，2017，37(6)：892-898.

[9] 葛朋，杨波，毛文彪，等.基于引导滤波的高动态红外图像增强处理算法[J].红外技术，2017，39(12)：1092-1097.

[10] 薛萍.基于超像素特征表示的图像前景背景分割算法[J].西安科技大学学报，2017，37(5)：731-735.

[11] 刘丽霞，李宝文，王阳萍，等.改进Canny边缘检测的遥感影像分割[J].计算机工程与应用，2019，55(12)：54-58.

[12] 田雯，胡玉荣.共生矩阵耦合Otsu阈值的彩色图像边缘提取算法[J].电子测量与仪器学报，2018，32(7)：52-60.

[13] 刘明纲.基于高斯混合模型和NSCT的图像边缘检测方法[J].科技通报，2018，34(8)：108-112.

[14] 陈石涛，杨龙兴，丁力，等.一种基于改进的SUSAN算法的焊缝边缘检测方法[J].热加工工艺，2018，47(15)：161-164.

[15] 许传祥，石青云，程民德.零对称和反对称二进小波及其在边缘检测中的应用[J].中国图象图形学报，1996(1)：4-11.

[16] LEE S，LEE D，PARK Y. Pupil segmentation using orientation fields，radial non-maximal suppression and elliptic approximation [J]. Advances in electrical and computer engineering，2019，19(2)：69-74.

作者简介：顾小平(1984—)，男，湖北武汉人，硕士，讲师，研究方向为数字媒体设计、产品设计。

(上接第152页)

[4] JEFFREY Donahue，LISA Anne Hendricks，SERGIO Guadarrama，et al. Long-term recurrent convolutional networks for visual recognition and description [C]// IEEE Conference on Computer Vision and Pattern Recognition. Boston：IEEE，2015：2625-2634.

[5] CHRISTOPH Feichtenhofer，AXEL Pinz，ANDREW Zisserman. Convolutional two-stream network fusion for video action recognition [C]// Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：204-215.

[6] ZHOU Bolei，AGATA Lapedriza，XIAO Jianxiong，et al. Learning deep features for scene recognition using places database [C]// Neural Information Processing Systems. Montreal：NIPS，2014：487-495.

[7] DIOGO C Luvizon，DAVID Picard，HEDI Tabia. 2D/3D pose estimation and action recognition using multitask deep learning [C]// 2018 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：5137-5146.

[8] 李军锋，何双伯，冯伟夏，等.基于改进CNN的增强现实变压器图像识别技术[J].现代电子技术，2018，41(7)：29-32.

[9] JOAO Carreira，ANDREW Zisserman. Quo vadis，action recognition? A new model and the kinetics dataset [C]// Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：6299-6308.

[10] 范晓杰，宣士斌，唐凤.基于Dropout卷积神经网络的行为识别[J].广西民族大学学报(自然科学版)，2017，23(1)：76-82.

[11] 周永生.基于LSTM神经网络的PM2.5预测[D].长沙：湖南大学，2018.

[12] WANG Limin，XIONG Yuanjun，WANG Zhe，et al. Temporal segment networks：towards good practices for deep action recognition [C]// European Conference on Computer Vision. Amsterdam：Springer，2016：20-36.

[13] WANG Huogen，WANG Pichao，SONG Zhanjie，et al. Large-scale multimodal gesture recognition using heterogeneous networks [C]// 2017 IEEE International Conference on Computer Vision Workshops. Venice：IEEE，2017：3129-3131.

作者简介：朱云鹏(1994—)，男，硕士研究生，主要研究方向为图像识别和机电一体化。

黄 希，男，副教授，主要研究方向为机电一体化、计算机辅助设计。