

基于时空双分支网络的人体动作识别研究

宫法明, 马玉辉

(中国石油大学(华东) 计算机科学与技术学院, 山东 青岛 266580)

摘要: 常规的人体动作识别算法在单一特定的场景中效果较为突出,但在海洋钻井平台的实际工程场景中,易受管道遮挡和干扰,不能充分地利用视频的时序结构信息。针对这些问题,提出了一种复杂场景下基于时空双分支网络的人体动作识别框架。采用多规则区域提案标记算法将海水区域分离,将先验知识加入支持向量机分类器,提出后验判别准则以去除非人员目标,通过目标定位与检测算法分割出人员目标,利用卷积姿态机算法进行身体部位定位和关联程度分析以提取全部人体关键点信息,形成关键点序列;通过双分支网络对人体关键点轨迹和光流轨迹叠加融合,完成了人体动作的分类与识别。实验结果表明,该方法实现了人体不可见关键点的检测和估计,免去了人工标注目标的繁杂工作,能够有效解决海洋平台场景下的人体动作识别问题。

关键词: 人体动作识别;关键点检测;目标检测;动作分类;卷积姿态机;深度学习

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2020)09-0023-06

doi:10.3969/j.issn.1673-629X.2020.09.005

Research on Human Action Recognition Based on Space-time Double-branch Network

GONG Fa-ming, MA Yu-hui

(School of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: The conventional human action recognition algorithm is more effective in a single specific scene. However, it is vulnerable to pipeline occlusion and interference in the actual engineering scene of offshore oil platforms, and the timing structure information of the video cannot be fully utilized. Aiming at these problems, a human action recognition framework based on spatiotemporal double-branch network in complex scenes is proposed. The multi-rule regional proposal marker algorithm is used to separate the seawater regions, the prior knowledge is added to the SVM classifier, and the posterior discriminant criterion is proposed to remove the non-personnel targets. The target location and detection algorithm is used to segment the human target, and the convolutional pose machines algorithm is used to perform body part localization and correlation degree analysis to extract all key information of the human body to form a key point sequence. The double-branch network superimposes the human key point trajectory and the optical flow trajectory, and completes the classification and recognition of human action. The experiment shows that the proposed method realizes the detection and estimation of the invisible key points of the human body, eliminating the complicated work of manually marking the target, which can effectively solve the problem of human action recognition under the ocean platform scene.

Key words: human action recognition; key point detection; object detection; action classification; convolutional pose machines; deep learning

0 引言

长期以来,因人体动作的复杂性和多样性,人体动作识别一直是研究人员重点研究方向。人体动作识别实际上是一个检测和分类问题,主要包括目标检测、人体关键点检测和定位^[1]以及动作分类与识别,近年来这一领域的研究取得了重大突破。在一些配备监控

设施的公共场所,人体动作识别的应用方便了公共场所的管理和安保工作。

文中以海洋钻井平台上的工作人员为研究对象。由于海洋平台具有远离陆地、救逃难度大等特点,在海上平台进行石油钻采生产作业的人员面临着各种危险,如坠海、设备坍塌等,所以对海洋平台状况的实时

收稿日期: 2019-11-04

修回日期: 2020-03-05

基金项目: 科技部创新方法工作专项资助项目(2015IM010300)

作者简介: 宫法明(1970-),男,教授,CCF会员(61798M),研究方向为计算机图形图像处理、大数据智能处理与云计算;马玉辉(1994-),男,硕士研究生,研究方向为图像处理、计算机视觉。

监控就显得尤为重要,这就要求在大量的监控视频中检测出人员目标并对其进行动作识别。以往的人体动作识别算法^[2]在单一特定场景下效果较突出,但在海洋钻井平台这种复杂场景中受背景变化的影响比较大,难以保证较高的识别准确率。此外,人体还可能被海洋平台上的密集管道所遮挡,由于管道颜色与工作人员的安全服颜色相近以及海水的流动和复杂天气的影响使得人体动作识别更加困难。针对海洋工作平台上的管道和设施是静止不动的这一特点,提出了一种多规则区域提案标记方法(multi-rule regional proposal marker, M-RPM)。该方法将海洋区域分离,在非海洋区域只检测运动物体,利用支持向量机(support vector machine, SVM)对形似人员的管道进行预判别,在此基础上检测出人员目标以减少计算的工作量。

人体动作识别往往与人体姿态估计^[3]紧密相连,人体姿态估计是对图像中人员目标的关键点进行识别和定位,深度卷积神经网络的推广,使人体姿态估计的问题得到进一步解决。人体姿态估计的方法主要分为两类:自上而下的方法和自下而上的方法。其中,自上而下的方法是指先检测到人员目标,然后使用目标包围盒^[4]进行定位,最后使用单人估计的方法定位人体的所有关节;自下而上的方法是指先定位到所有关节的位置,再区分关节的从属目标,最后将关节组装成一个完整的人体姿态。前者适用于人员目标稀疏的情况,后者适用于人员目标密集的情况。针对海洋平台这个特殊场景,提出改进的卷积姿态机算法(convolutional pose machines, CPM)实现人体关键点的检测,通过目标检测定位出海洋平台上的人员位置,由于目标检测存在一定的误差,采取多尺度的方式得到完整的人员目标,以便在人体关键点检测阶段提取到人体所有关键点坐标,提高人体动作识别的准确率。最后,将人体关键点坐标序列形成动作轨迹记为空间信息,光流轨迹为时间信息,两者进行融合,实现人体动作的分类和识别。

文中的主要贡献有:在海洋平台这个复杂场景中,为了降低海水流动对目标检测的影响,提出一种基于M-RPM的数据预处理方案,提高目标识别的准确率;为了降低柱形管道对正样本的影响,将先验知识加入目标检测,提出SVM的后验判别;在有遮蔽物的情况下提出改进的CPM算法,利用目标检测的结果实现人体不可见关键点的检测和估计,免去人工标注目标的繁杂工作。

1 相关工作

Alexander 等人^[5]提出了结合卷积神经网络和级联的方法,通过初步计算得到一个节点的坐标,然后根

据该坐标在原始图像中获得对应的局部图像,利用该局部图像完成更高精度的坐标计算。该方法对于分辨率较低的原始图像效果较差,同时由于采用了级联的方法,每一个节点的坐标需要进行重复的卷积操作,计算复杂度高。为了解决上述问题,Varun 等人^[6]提出了一个基于 CPM 的框架,应用于人体姿态估计,使用卷积层表达纹理信息和空间信息,通过人体各部位的响应图来表达各部位之间的空间约束关系。在同一个网络中,多尺度处理输入的特征图和响应图,既能确保精度,又考虑了各部位之间的距离关系。

基于静态单帧的人体姿态估计只依赖于空间信息难以解决人体部位遮挡和人体连续动作识别问题^[7],对于部位遮挡问题,旷视科技 Face++^[8]提出了一种新的网络结构,即为 CPN 级联金字塔网络,该算法包含 GlobalNet 和 RefineNet 两个阶段。GlobalNet 是一个特征金字塔网络,它能够获得所有简单的关键点,RefineNet 是专门用来处理重叠不可见的关键点,通过逐步细化的流水线将所有级别的特征表征和关键点的挖掘损失集成到一起,该算法在 COCO 的关键点检测比赛中取得了最佳成绩。为了更好地学习时空特征,文献[9]提出了一种视频识别的双流体系结构,通过 ConvNet 传递空间信息即单个静态 RGB 帧和另一个时间信息即多个帧的光流,然后融合两个流的并行输出,形成最终的分值,实现了基于视频的人体动作识别。综上所述,近几年在人体动作识别的研究取得了显著的科研成果,但在如海洋采油平台这种复杂环境下人体动作识别的效果亟待提高。

2 人体动作识别框架

文中提出的框架以视频信息作为输入,分别生成单帧静态图和多帧光流图,将 RGB 静态图像输入到空间分支网络对目标检测与分类进行预处理,提取空间维度上的高层特征以生成初始像素级的标记,利用空间分支网络生成二进制对象分割图像信息。将连续的光流图序列输入时间分支网络以利用时序结构信息,将空间外观图映射到视频帧前景图上以计算每一帧的二进制对象分割。然后,进入时空双分支网络训练的目标检测器,判断是否存在目标对象以及检测出目标对象可能存在的区域,对区域候选边界框和对象真实边界框之间的重叠度进行评分,得到目标包围盒和坐标信息,为人体关键点提取提供数据来源。文中将目标检测和基于改进的 CPM 算法充分融合,实现人员目标检测和关键点提取一体化,将连续帧的关键点坐标信息构成动作轨迹,视为空间流信息。此外,通过提取多帧光流信息将连续帧之间的光流堆叠形成光流轨迹,视为单位时间内位移信息。最后,对动作轨迹和光

流轨迹融合叠加,实现人体动作分类与识别,框架的流程如图1所示。

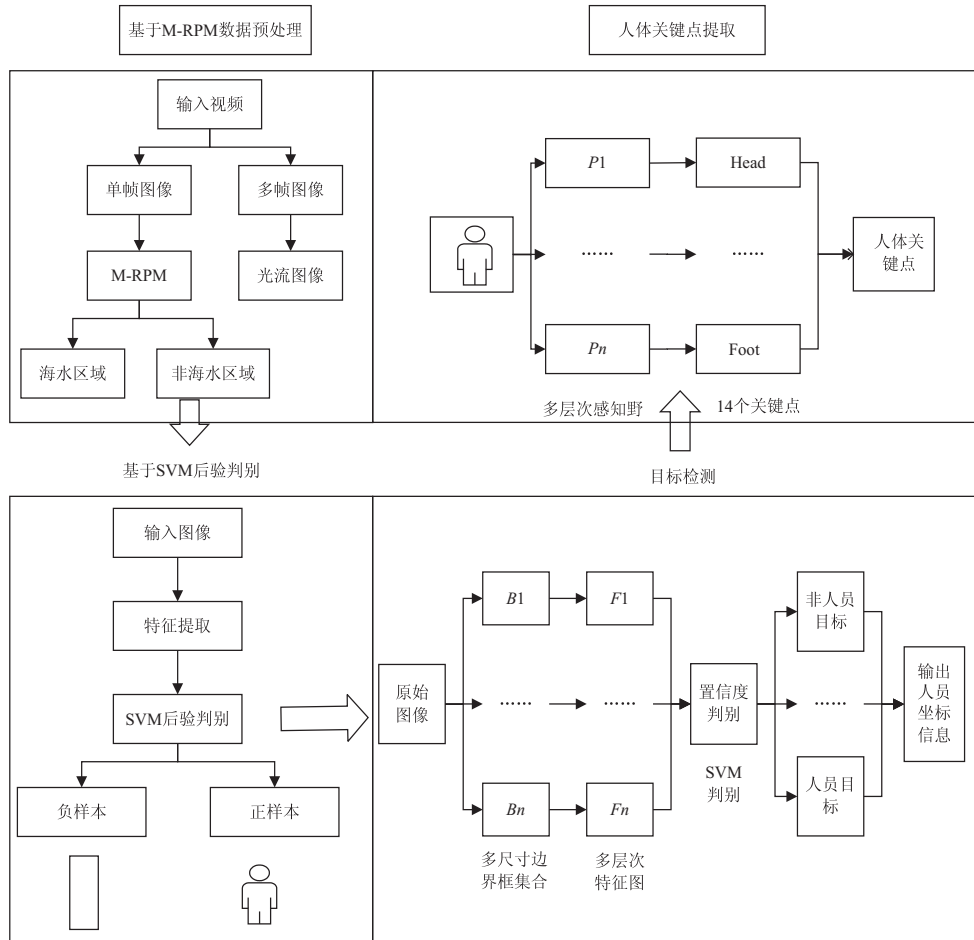


图1 人体动作识别框架流程

2.1 基于 M-RPM 的数据预处理

海洋平台监控中心获取到的监控视频中会存在大量的海洋区域,海水的流动会影响目标检测的准确性。针对此问题,提出了一种基于 M-RPM 的数据预处理方法。

根据场景的不同定义适合此场景的方案,使用一种过分割手段将图像分割成若干小区域,对不同的区域依次编号,起始序号为 x_0 。通过选择性搜索和合并规则^[10]生成两个可能性最高的区域方案,重复此过程直到整张图像合并成两个界限相对明显的区域,即为海洋区域和非海洋区域。

图像区域划分的过程中优先合并以下四种区域^[11]:纹理相近区域、颜色相近区域、合并后区域总面积最小区域以及合并后总面积在其原始图像中所占比例最大区域。前两者分别通过梯度直方图和颜色直方图衡量,面积最小保留规则保证合并操作的尺度均匀,避免出现大区域连续吞并其他小区域的现象,占空比最大原则保证区域合并后形状规则、易于区分。通过颜色、纹理、面积和位置生成的区域特征可以直接由子区域特征计算得到,计算速度较快。此外,区域标记也

是对后续目标识别结果的预判规则,具体生成过程如下:

首先,设有 8 个区域标记为 x_0 到 x_7 ,各区域面积相等,选用面积最小保留规则,依据两区域最临近规则,如以下公式所示,依次合并得到区域 $Q_1(X)$;

$$Q(X) = \sum_{n=0}^{n \leq 4} \sum_{i=2n} (X_i \cup X_{i+1})$$

然后,当 $n=4$ 时,依次迭代合并 $2n$ 次,得到合并后的区域 $Q_2(X)$ 。若在合并过程中出现吞并现象,区域合并时重复以上过程,执行 i 次迭代,得到与上述结果相同的区域划分,但是迭代次数明显增加。在选用面积最小保留规则进行区域合并时应避免上述情况的发生,因为划分的小区域越多,迭代次数会不断增加,最终陷入无限迭代循环,导致区域标记失败。

通过上述过程,计算获得了描述图像区域的标记方案,为尽可能不遗漏所有区域,上述操作在多个颜色空间中同时进行,如 RGB、HSV 和 Lab 等。在同一个颜色空间中使用上述四条规则的不同组合进行合并,所有颜色空间与所有规则的全部结果在去除重复区域后都作为区域标记的输出,最终得到海洋区域与非海洋区域的明显界限。

2.2 目标检测器

对于单帧静态图像上的人员目标检测,文中在单发多盒探测器^[12] (single shot multibox detector, SSD) 上做出改进,改进算法的主要步骤如图 2 所示。

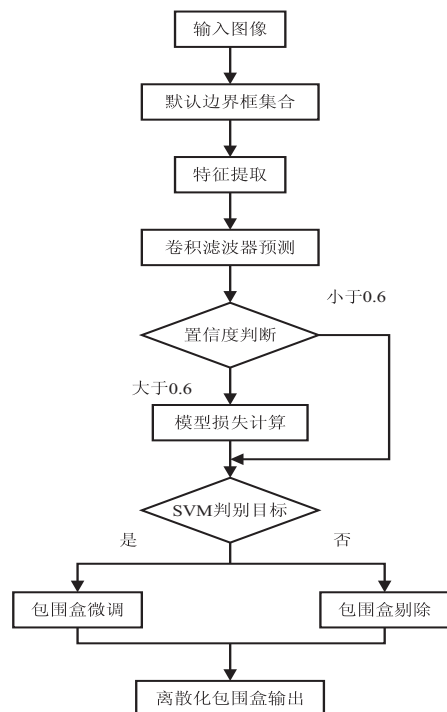


图2 目标检测算法流程

Step1: 对于不同尺寸的输入图像产生一组固定大小的默认边界框集合;

Step2: 对该组默认边界框内的区域进行特征提取,针对本场景下工作人员的形体表征,以颜色、形状和纹理等为主要特征进行提取以形成不同层次的特征图单元;

Step3: 将每个层次的特征图单元以卷积的方式平铺特征映射,使得每个默认边界框与相对应的特征图单元的位置是固定的;

Step4: 在每个特征图单元上使用小卷积核卷积滤波器^[13] 预测每一个边界框中的物体并计算出置信度;

Step5: 将实际置信度与预设置置信度进行判别,对于大于置信度阈值 0.6 的情况进行模型损失的计算;

Step6: 通过 SVM 后验判别,在人员目标的先验知识基础上实现精准目标检测,若判别为人员目标,则需对目标包围盒使用线性回归器进行微调处理,精细修正包围盒位置,否则视为无效包围盒,进行剔除操作;

Step7: 输出一系列在不同层次上的离散化目标包围盒,且具有不同的长宽比尺度。

在置信度判别过程中,每组默认边界框需计算出与相对应的实际边界框的误差和相应的评分,然后预测区域内的所有目标的类别和置信度,置信度阈值^[14] 设置为 0.6,即大于该阈值的对象类别视为目标类别。

置信度判别是目标检测的初筛选过程,将默认边界框与任何具有高于阈值的实际边界框进行重叠度匹配,通过 SVM 后验判别简化了匹配过程。此外,本算法允许预测多个重叠的默认边界框的评分,而不是只挑选具有最大重叠度的边界框进行评分预估。

2.3 基于改进 CPM 的人体关键点提取

CPM 的每个阶段为人体每个部位重复生成 2D 置信图^[15],该置信图与图像特征同时用作下一阶段的输入,为人体各关键点的位置提供更精确的估计。文中在 CPM 的基础上对输入图像做出改进,将上一阶段目标检测得到的离散化人员目标包围盒坐标作为改进后 CPM 的原始输入,免去人工标注目标的繁琐工作,实现了人体关键点的自动定位与识别。

该算法以上一阶段目标检测得到的 $w * h$ 大小的彩色图像作为输入,采取多尺度的方式,按照 1.0 比 1.2 倍的比例扩大感知野。经过 VGG 的前 12 层网络的特征提取得到一个特征映射 F ; 通过关键点分析结果,算法分成两个循环分支,一个分支用于预测身体部位位置的二维置信度图 S ,进行身体部位定位和关键点检测得到人体所有可见的关键点,另一个分支用于预测像素点在骨架中的二维矢量场 L ,进行关联程度分析和部位亲和力计算得到人体不可见关键点的信息;当前阶段 $t(t \leq p)$ 时,循环分支以特征图 F 作为输入,得到一组 S_t, L_t ; 之后的分支分别以上一个分支的输出 S_{t-1}, L_{t-1} 和特征图 F 作为输入,不断进行迭代;经过 p 个阶段最终输出 $S(p)$ 和 $L(p)$ 计算 S, L 的预测值与 ground truth 之间的 L2 范数, S 和 L 的 ground truth 根据标注的 2D 点计算,如果某个关键点标注缺失,则不计算该点的值,最终输出所有关键点的信息。

2.4 人体动作分类与识别

对人体的动作进行分析时更多关注的是局部细节动作,但在视频监控中细节动作特征往往表现得并不明显。通过层次化处理人体关键点坐标得到粗分类动作^[16],在此基础上完成动作识别任务,这种方式也具有较好的识别能力。通过判断人体部位关键点位置变化的缓慢程度,将动作粗分类为头部动作、上肢动作、躯干动作和下肢动作。对于不同类别的动作,轨迹关注点亦不相同。对于上肢和下肢动作,主要关注手部和腿部的关键点轨迹变化,而对于躯干动作,往往关注身体中心的关键点轨迹变化。通过改进的 CPM 算法得到每组粗分类动作的关键点序列,完成人体动作的分类。

对于局部细节动作的识别,用粗分类动作的关键点序列表示动作轨迹,通过叠加多帧光流得到密集光流轨迹。文中根据两个不同的识别流从空间和时间的角度通过连接各个局部动作片段的特征描述整个动作

序列^[17]。空间流在单帧静态图像上将每个轨迹点映射到人体关键点上,时间流以密集光流的形式从运动中识别动作,利用动作轨迹和光流轨迹的叠加作为动作信息。前者考虑整个序列中每个点的位移,而后者侧重于连续帧之间每个点的位移。最后,通过比较两轨迹间的相似性,完成动作分类和识别任务。

3 实验仿真与分析

3.1 实验数据集

原始数据来自于海洋采油厂的流媒体服务器。在原始视频库数据集上,使用关键帧图像提取法选取带有目标的图像数据集,即在1秒的间隔内将首帧、中间帧和尾帧视为关键帧图像,形成目标检测所使用的标签数据库。该数据库存储了目标的标签类型和位置信息,包含了3万张目标图像,由424路摄像头采集各个场景的图像组成。人体关键点检测形成的点集数据库存储了关键点序列,包括图像的名称、人体14个关键点以及关键点的坐标序列。

针对海洋平台这个特殊场景,文中从安保工作的角度预置了6种动作类型,包括打电话、跌倒、跑步、站立、弯腰和行走,采集动作序列数据作为人体动作模型库标准。其中,由于弯腰和跑步发生的时间间隔较为短暂,摄像头难以捕捉采集到大量的数据。此外,跌倒和打电话行为不经常出现,采集到的数据也较少,每种动作类型具体的数据分布如图3所示。

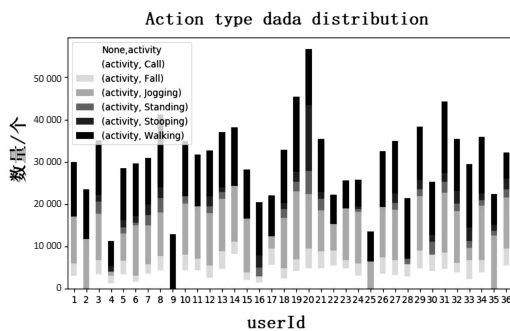


图3 不同动作类型的数据分布

3.2 实验设计与分析

在目标检测实验中,由于管道颜色与工作人员的安全服颜色相近以及柱形管道形状的影响,使得人员目标的误报率上升。针对这一问题,文中采用SVM算法预先训练目标分类器,使得在海洋平台场景下能够准确地区分出人员和管道目标。然后,将该分类器与目标检测的模型相结合,减少由于柱形管道而产生的误报率。目标分类器的使用极大提高了目标检测的准确率。

为了验证该方法的有效性,分别选取Faster-RCNN^[18]、MobileNet-SSD^[19]以及SSD^[12]算法进行对

比实验,如表1所示。

表1 目标检测各方法在海洋平台场景下的准确率%

数量(张)	RCNN	M-SSD	SSD300	文中方法
5 000	63.2	64.8	69.5	73.7
10 000	69.3	68.9	72.1	79.6
15 000	75.85	74.2	78.7	84.5
20 000	80.1	79.3	82.3	87.4

实验证明:在相同数据规模的验证实验中,提出算法的检测准确性优于三组对比实验,且随着数据集规模的扩大,准确性要远高于对比实验,从而得出结论,提出的目标识别算法对于复杂场景下的目标检测有着显著的准确率。改进后的CPM算法按照从头部开始自上而下的顺序依次检测人体各部位,将人体分为头部、颈部、右肩、右肘、右手腕、左肩、左肘、左手腕等14个部位^[20]。通过对每个部位位置的预测计算出置信度,产生人体各部位的响应图。根据人体关键点序列,得到的人体关键点部位分布,最终构建出人体的2D关节图^[21],如图4所示。



图4 基于2D关键点的人体动作识别示意图

通过选取基于模板匹配算法进行对比实验,结果表明,所提出的框架对6种基础动作的识别率要具有普适性。其中,对于站立和跌倒动作,主要检测躯干中心的动作轨迹^[22],归类为躯干动作;对于行走,主要检测下肢动作和整体部位的运动位移,归类为下肢动作;对于打电话,主要考虑上肢的运动轨迹,将其归类为上肢动作。由图5中数据可得:站立和跌倒动作识别结果较好,由于行走时受上肢的摆动的影响以及打电话时头部的晃动,后两个动作识别准确率还有待提升。

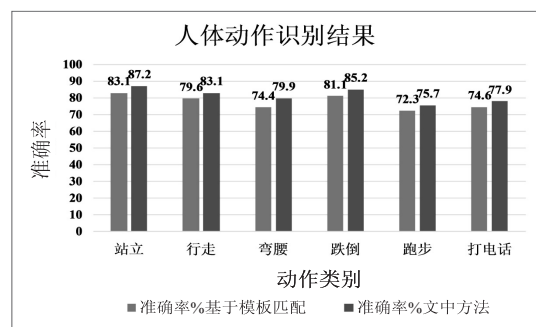


图5 海洋平台场景下的人体动作识别结果

综合上述实验结果,表明所提出的基于复杂场景下的人体动作识别算法能够准确识别出人员目标,并且能够准确估计出被遮挡的部位,从而得到完整的人体关键点序列,基本能够满足人体动作识别的准确性要求。

4 结束语

结合目标检测与人体关键点提取等技术,实现了复杂场景下的人体的动作识别。通过 M-RPM 算法降低海水流动和复杂石油管道对人员目标检测造成的误差,在多种传统人体动作识别算法的基础上进行有效融合,使之适用于远离陆地的海洋平台环境,从而保证工作人员的安全以及平台工作的顺利开展。尽管提出的方法在目标检测和人体关键点识别方面取得了较好结果,但对局部细微动作的识别还有待提高。因数据集规模有限,未能更多地考虑复杂动作等潜在问题。如何解决这些复杂动作的识别问题将会是下一步的主要研究工作。

参考文献:

- [1] CHEN Y, SHEN C, WEI X, et al. Adversarial PoseNet: a structure-aware convolutional network for human pose estimation[C]//International conference on computer vision. Venice, Italy: IEEE, 2017: 1221-1230.
- [2] 秦磊, 胡琼, 黄庆明, 等. 基于特征点轨迹的动作识别[J]. 计算机学报, 2014, 37(6): 1281-1288.
- [3] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3d human pose estimation[C]//International conference on computer vision. Venice, Italy: IEEE, 2017: 2659-2668.
- [4] CHERON G, LAPTEV I, SCHMID C, et al. P-CNN: pose-based CNN features for action recognition[C]//International conference on computer vision. Santiago, Chile: IEEE, 2015: 3218-3226.
- [5] TOSHEV A, SZEGEDY C. DeepPose: human pose estimation via deep neural networks[C]//Computer vision and pattern recognition. Columbus, OH: IEEE, 2014: 1653-1660.
- [6] WEI S, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]//Computer vision and pattern recognition. Las Vegas: IEEE, 2016: 4724-4732.
- [7] CAO Z, SIMON T, WEI S, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Computer vision and pattern recognition. Washington: IEEE, 2017: 1302-1310.
- [8] CHEN Y, WANG Z, PENG Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Computer vision and pattern recognition. Utah: IEEE, 2018: 7103-7112.
- [9] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 1(4): 568-576.
- [10] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Computer vision and pattern recognition. Ohio: IEEE, 2014: 580-587.
- [11] 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.
- [12] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European conference on computer vision. Amsterdam, Netherlands: IEEE, 2016: 21-37.
- [13] 王守义, 周海英, 杨阳. 基于卷积特征的核相关自适应目标跟踪[J]. 中国图象图形学报, 2017, 22(9): 1230-1239.
- [14] CARREIRA J, AGRAWAL P, FRAGKIADAKI K, et al. Human pose estimation with iterative error feedback[C]//Computer vision and pattern recognition. Las Vegas, Nevada: IEEE, 2016: 4733-4742.
- [15] LI Y, LAN C, XING J, et al. Online Human action detection using joint classification-regression recurrent neural networks[C]//European conference on computer vision. Amsterdam, Netherlands: IEEE, 2016: 203-220.
- [16] 尹建芹, 刘小丽, 田国会, 等. 基于关键点序列的人体动作识别[J]. 机器人, 2016, 38(2): 200-207.
- [17] NEWELL A, YANG K, DENG J, et al. Stacked hourglass networks for human pose estimation[C]//European conference on computer vision. Amsterdam, Netherlands: Springer International Publishing, 2016: 483-499.
- [18] PENG X, SCHMID C. Multi-region two-stream R-CNN for action detection[C]//European conference on computer vision. Amsterdam, Netherlands: Springer International Publishing, 2016: 744-759.
- [19] HOWARD A, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[C]//Computer vision and pattern recognition. Washington: IEEE, 2017.
- [20] BELAGIANNIS V, AMIN S, ANDRILUKA M, et al. 3D pictorial structures for multiple human pose estimation[C]//Computer vision and pattern recognition. Ohio: IEEE, 2014: 1669-1676.
- [21] PISHCHULIN L, INSAFUTDINOV E, TANG S, et al. DeepCut: joint subset partition and labeling for multi person pose estimation[C]//Computer vision and pattern recognition. Las Vegas, Nevada: IEEE, 2016: 4929-4937.
- [22] 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. Las Vegas, Nevada: IEEE, 2016: 1010-1019.