基于单目视觉的在线人体康复动作识别

闫 航[1,2] 陈 刚[2*] 崔莉亚[1,2] 张乐芸[3] 胡北辰[1]

[1](郑州大学信息工程学院 河南 郑州 450001)

[2](郑州大学互联网医疗与健康服务协同创新中心 河南 郑州 450052)

[3](郑州大学护理与健康学院 河南 郑州 450001)

摘 要 为了进一步提高居家监护场景下人体动作识别的可靠性与实时性,更好地辅助出院后的卒中患者进行康复训练,提出一种基于单目视觉的在线人体动作识别算法。融合姿态估计 OpenPose 与最近邻匹配算法对监控视频流中的目标人体生成动作序列。通过滑动窗口选取原始姿态特征并对其预处理转化为鲁棒性特征,输入到多层 LSTM 长短时记忆网络中进行康复动作识别。实验结果表明,该方法对活动背景、人体穿着、无关人员的干扰等具有较强的适应能力,能够在线识别连续的康复动作且准确率达 90.66%,在居家康复训练场景中有一定的应用价值。

关键词 姿态估计 动作识别 长短时记忆网络 康复训练 居家看护

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.02.029

ONLINE HUMAN REHABILITATION ACTION RECOGNITION BASED ON MONOCULAR VISION

Yan Hang[1,2] Chen Gang[2*] Cui Liya[1,2] Zhang Leyun[3] Hu Beichen[1]

[1](*College of Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China*)

[2](*Internet Medical and Health Service Collaborative Innovation Center, Zhengzhou University, Zhengzhou 450052, Henan, China*)

[3](*College of Nursing and Health, Zhengzhou University, Zhengzhou 450001, Henan, China*)

Abstract In order to further improve the stability and real-time of human action recognition in the home monitoring scene, and better assist discharged stroke patients to perform rehabilitation training, we propose an online action recognition algorithm based on monocular vision. OpenPose and nearest neighbor matching algorithm were combined to generate action sequences for target human body in surveillance video streams. The original features were selected by sliding window and transformed into robustness features. The features were input into multi-layer LSTM for rehabilitation action recognition. The experimental results show that the proposed method has strong adaptability to the background of activities, human body wear, interference from unrelated personnel, etc. It can recognize continuous rehabilitation actions online with an accuracy rate of 90.66%, which has certain application value in home rehabilitation scene.

Keywords Pose estimation Action recognition LSTM Rehabilitation training Home care

0 引 言

近年来基于视觉的人体动作识别的研究得到了广泛的关注,是计算机视觉中一个极具挑战性的课题,其涉及图像处理、模式识别、人工智能等多个学科,在智能监控、人机交互、康复运动等领域有着广泛的应用前景[1]。脑卒中是最常见的慢性病之一,具有高发病率、高致残率的特点,是老年人健康的重大威胁。而康复锻炼是恢复卒中患者日常生活能力的主要手段,也是

收稿日期:2019-08-09。国家重点研发计划项目(2017YFB1401200);河南省科技攻关项目(182102310137);赛尔网络下一代互联网技术创新项目(NGII20170716);2019 年河南省高等学校重点科研项目(19A520039)。**闫航**,硕士生,主研领域:计算机视觉,动作识别。**陈刚**,工程师。**崔莉亚**,硕士生。**张乐芸**,硕士生。**胡北辰**,硕士生。

广泛推荐的康复疗法[2]。当前在居家康复领域缺乏护理医师的现场指导,同时存在看护者缺乏耐心和信心、康复知识不足的问题,导致患者出院后在家中难以完成有针对性的康复目标,依从性较差[3]。因此建立一种居家康复场景下的在线动作识别模型,实现患者康复过程中动作的实时监督与指导,对患者中长期的康复水平有着重要的意义。

根据获取数据的方式,可以分为基于传感器和基于视觉的人体动作识别。基于传感器的动作识别方式起步较早,且在康复动作识别领域中也有较多的研究。Bisio 等[4]采用三轴加速度计采集病人运动信息,通过 SVM 分类器对手臂伸展、肩关节屈伸等康复动作取得了良好的识别效果。马高远等[5]对采集的上肢肌电信号通过小波分解提取特征,在 8 种常用康复动作上取得了 92.86% 的识别率。复杂动作通常需要多个传感器协同工作才能达到较好的识别效果,然而该方式会给身体带来极大的不适。基于视觉的动作识别主要分为人工特征和深度学习特征两类。传统的采用人工特征的动作识别方法侧重于局部特征提取,Wang 等[6]提出改进的稠密轨迹用于动作识别,提升了对复杂场景的鲁棒性。深度学习能自主提取具有强大表征能力的特征,逐渐获得了更多的关注。主流的深度学习模型有 3D CNN[7]、Two-Stream CNN[8]、LRCN(Long-term Recurrent Convolutional Network)[9]、R-C3D(Region Convolutional 3D Network)[10]等。然而以上模型对整幅视频帧进行深层卷积操作,存在复杂度高、运算速度慢、训练困难等问题,制约了其在现实生活中的应用。基于视觉的康复动作识别研究工作较少且主要采用骨架特征。人体骨架特征包含的运动信息比较完整,对于肢体动作来说是一种良好的表征方式[11]。邵阳等[12]采用深度相机 Kinect 获取人体骨架信息,通过基于余弦的动态时间规整方法有效识别了 6 种上肢训练动作。唐心宇等[13]提出一种基于 Kinect 的三维人体姿态估计方法,并通过计算关节角度来评估指定的康复动作。文献[14]同样对 Kinect 获取的三维人体骨架关键点进行角度计算,与标准的动作规范进行比较来识别动作并指导康复训练。以上康复动作识别方法存在如下问题:每个动作需要人工建立复杂的对照模型,不易拓展且泛化能力较差,对人员位置、角度均有严格的要求;缺乏对多人场景的兼容问题,居家康复环境下易受干扰;相比于单目摄像头,深度相机 Kinect 较为昂贵、普及度不高且设备难以获取;多数算法仅能处理经过裁剪的视频段或识别过程需要繁琐的人工干预,难以实现在线、连续的动作识别。

在线动作识别更具挑战性:预定义的目标动作发生时间不确定;除了目标动作外还存在其他动作与状态;处理速度能够匹配上监控视频流。在线动作识别对于应用落地具有重要的现实意义,但是相关的研究工作却很少。Li 等[15]基于 Kinect 获取 3D 骨架数据,提出了一种联合分类和回归的 LSTM 网络实现了单人场景下的在线动作识别,但其识别的是持续时间较短的日常活动。

当前大多数基于视觉的人体动作识别算法存在复杂度高、建模困难、无法处理在线视频、部署条件苛刻等问题,而基于传感器的动作识别会对人体造成极大的不适。为了更好地适用于居家场景下卒中患者的康复动作识别,本文设计并实现了一种基于单目视觉的在线动作识别算法。采用姿态估计算法 OpenPose 对单目摄像头获取的连续视频流进行骨架关键点提取并结合最近邻匹配生成目标动作序列,对人体多个关节的运动变化进行充分的表征,同时避免了其他图像区域带来的噪声干扰。通过实验选取合适的滑动窗口大小,在目标人体的动作序列上通过滑动窗口提取原始特征并进一步预处理为鲁棒性特征,输入到预训练的 LSTM 分类网络中进行康复动作识别。本文提出的方法在康复训练场景中能够有效进行在线动作识别,模型易于部署,一定程度上能够适应非理想、嘈杂的环境,对于脑卒中患者的康复具有重要的意义。

1 相关算法

1.1 姿态估计 OpenPose

姿态估计 OpenPose[16]是首个基于深度学习实现的实时多人姿态估计开源库,能够实时地对图片中每个人的姿态进行精准的估计,实现面部、躯干、四肢、手部骨骼点的提取。它兼顾了实时性与准确性,且具有较强的鲁棒性。

该方法的核心是一种利用 Part Affinity Fields(PAFs)的自下而上的人体姿态估计算法,即先检测关键点再获得骨架,在多人的场景下避免了过多的运算时间。图 1 所示为 OpenPose 的多阶段预测网络结构,该框架以 VGG-19 网络模型为基础,将输入的图像转化为图像特征 F,通过分阶段预测分别回归 $L(p)$ 与 $S(p)$,其中:$L(p)$ 表示亲和度向量场 PAFs,描述关键点在骨架中的走向;$S(p)$ 表示关键点的置信度。该结构将每一次预测分为 6 个阶段,通过前 4 个阶段预测亲和度向量场 L^t,后 2 个阶段预测置信度 S^t。在每个后续阶段,将前一阶段的预测结果与原始图像特征连接起来作为输入,用于生成更精细的预测。在得到关键

点的置信度及亲和度之后,采用匈牙利算法对相邻关键点进行最优匹配,从而得到每一个人的骨架信息。OpenPose 的实时性非常出色,采用单目摄像头即可获取可靠的关键点信息,无需专用的深度摄像头。

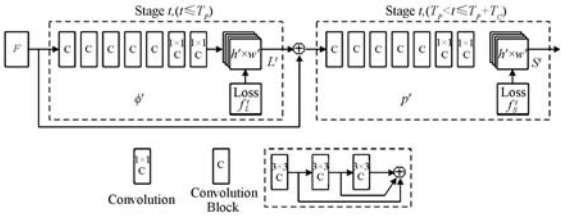

图 1　多阶段预测网络架构

1.2　循环神经网络

循环神经网络(Recurrent Neural Network,RNN)是一种利用上下文状态来挖掘数据中时序信息的深度神经网络。相比于卷积神经网络,RNN 会对于每一个时刻的输入结合当前模型的状态计算输出。单纯的 RNN 存在长期依赖问题,可能会丧失学习远距离信息的能力。长短时记忆网络(LSTM)的出现成功解决了梯度消失问题,是当前最为流行的 RNN 网络,广泛应用于语音识别、自然语言处理、视频描述、动作识别等领域。

图 2 为 LSTM 的网络结构示意图,LSTM 的输入包括当前时刻网络的输入 $\boldsymbol{x}_t$、上一刻 LSTM 的输出 $\boldsymbol{h}_{t-1}$、上一时刻的记忆单元 $\boldsymbol{c}_{t-1}$,输出包括当前时刻的输出 $\boldsymbol{h}_t$ 和当前时刻的记忆单元 $\boldsymbol{c}_t$。

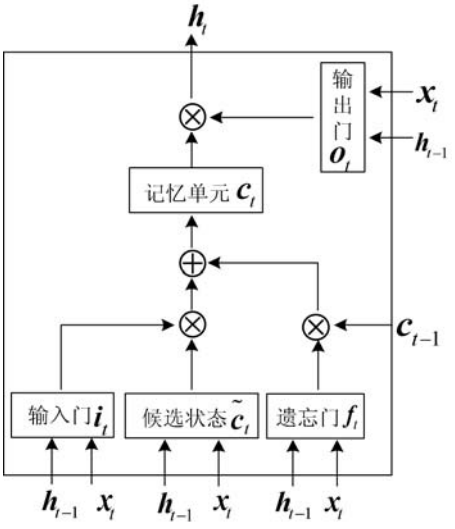

图 2　LSTM 网络结构示意图

LSTM 通过输入门与遗忘门控制记忆单元并结合输出门从而更有效地刻画长距离依赖。输入门、遗忘门与输出门的计算如下:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \tag{1}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \tag{2}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \tag{3}$$

式中:$\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_o$ 分别为输入门、遗忘门和输出门的权重矩阵;$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$ 分别为输入门、遗忘门和输出门的偏置。LSTM 的输出由记忆单元与输出门联合计算如下:

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_c \cdot \boldsymbol{x}_t + \boldsymbol{b}_c) \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \times \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \times \tilde{\boldsymbol{c}}_t \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \times \tanh(\boldsymbol{c}_t) \tag{6}$$

式中:$\tilde{\boldsymbol{c}}_t$ 是 t 时刻的候选状态;$\boldsymbol{W}_c$ 为候选状态的权重矩阵;$\boldsymbol{b}_c$ 是候选状态的偏置;$\boldsymbol{c}_t$ 为 t 时刻的记忆单元;$\boldsymbol{h}_t$ 则为 t 时刻最终的输出。

2　在线动作识别算法

2.1　算法框架

本文提出的在线动作识别算法主要由动作信息采集、特征提取和分类网络组成。算法的识别框架如图 3 所示,其输入为单目摄像头获取的实时监控视频流,首先采用姿态估计 OpenPose 提取图像帧中的多人骨架关键点,结合最近邻匹配算法在持续的监控流中生成目标人体动作序列,记录人体手部、手臂、腿部、颈部等多个关节的运动轨迹。动作序列为具备时序关系的连续 2D 骨架关键点,每帧提取的骨架信息作为一个时间步长。在动作序列上通过滑窗选取原始骨架关键点特征,经过坐标归一化、绝对坐标转相对坐标从而转化为鲁棒性特征并输入到构建的 LSTM 分类网络中,通过 Softmax 分类器判断三种康复动作并区分正常活动所表现的动作与状态。

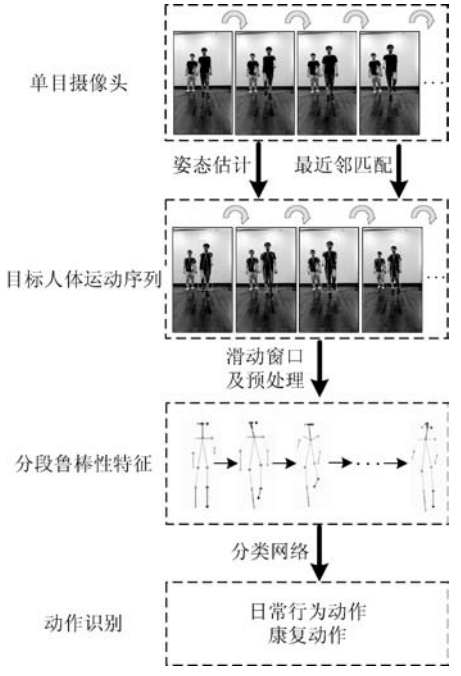

图 3　在线动作识别算法框架

滑窗大小的选择通过实验并结合平台处理速度与动作持续时间,LSTM 分类网络则采用经过裁剪的康

复动作数据集进行训练,将训练好的网络参数迁移到在线动作识别算法。该算法通过2D骨架关键点的时序信息进行动作识别,相比于双流、3D CNN等算法具有极强的速度优势。由于每一帧只对提取的18个人体骨架关键点进行处理,构建的LSTM网络相对于主流方法中的CNN网络而言其参数大大减少,模型易于优化从而避免了对海量数据集的依赖。

2.2 算法流程

2.2.1 动作信息采集

本文以1 920×1 080分辨率的单目摄像头获取实时监控视频流,用于在线动作识别。采用智能手机以及单目摄像头采集经过裁剪的康复动作数据,用于训练分类网络。为了提高识别算法的鲁棒性和给予被监护人员一定的自由度,数据集存在角度、远近、背景、分辨率的差异。

利用姿态估计方法OpenPose提取骨架关键点,通过VGG-19网络将输入的图像转化为图像特征 F,然后通过多层CNN分别预测关键点置信度与亲和度向量,联合置信度与亲和度向量得出人体的骨架信息。训练过程中模型总损失为置信度网络与亲和度向量场网络两者的损失之和,通过不断迭代完成神经网络参数的更新。由于姿态估计模型需要大量标注人体关键点的数据集来训练,为了达到更准确的效果,采用在超大规模图像数据集COCO中预训练的参数来初始化网络。将图像的分辨率调整为432×368后输入到模型中,输出为人体的18个2D骨架关键点,包括左右耳、左右眼、鼻、脖、左右肩、左右肘、左右腕、左右胯、左右膝和左右脚踝。图4所示为视频流中人体18个骨架关键点的检测结果,展示了“慢走”动作发生过程中人体关键点的变化。

图4 视频流中的人体骨架关键点检测

裁剪的视频段为单人视频且只发生一种动作,根据时序关系以一定的间隔对整个视频采样图像并提取骨架关键点来训练分类网络。视频流中提取的骨架信息仍是单帧的图像中独立的检测结果,目标人体在多人场景下失去时序关系,对于监控视频流则结合最近邻匹配算法生成目标人体动作序列。所设计的方法步骤如下:

1)确认目标人体。摄像头开启后通过OpenPose实时提取图像中的多人骨架关键点,计算每个人体 i 关键点的y轴坐标最大值、最小值之差 $y_{\mathrm{d}}^{i}=y_{\max}^{i}-y_{\min}^{i}$,$y_{\mathrm{d}}^{i}$ 值最大的人体 i 则认为距离摄像头最近,确定为在线动作识别的目标人体并为其创建一个动作序列。

2)目标最近邻匹配。以脖子部位的关键点坐标为基准坐标,计算当前帧每个人体 i 的基准坐标 (x_1^i, y_1^i) 与前一帧目标人体基准坐标 (x_0, y_0) 的欧氏距离 d_i,d_i 最小者判断为目标人体。d_i 计算方法如下:

$$d_i=\sqrt{(x_0-x_1^i)^2+(y_0-y_1^i)^2} \tag{7}$$

3)生成动作序列。结合最近邻匹配结果将视频流中目标人体的18个骨架关键点按照时序关系加入到动作序列中,若连续10帧没有检测到目标人体,则删除目标动作序列,重新执行步骤1)以确认目标人体。

2.2.2 特征提取

本文通过滑动窗口的方式从动作序列中提取原始特征,每帧的目标人体有18个2D骨架关键点,共36个特征。设置滑动窗口的大小为 n,即连续 n 帧图像作为一组分段特征,滑窗间隔设置为 k 帧。为了合理利用资源,设计队列的方式进行滑窗处理,假设动作队列为 T,滑窗提取分段特征的流程如下:

1)目标人体的骨架关键点不断加入队列直到队列长度为 n,即 $T=[T_1, T_2, \cdots, T_n]$,提取该分段特征。

2)从队头删除 $T_1, T_2, \cdots, T_K$,队尾不断加入后续的 k 帧关键点即 $T_{n+1}, T_{n+2}, \cdots, T_{n+k}$,提取该分段特征。

3)重复步骤2),直至该目标消失。

上述提取的分段特征仍是原始的骨架关键点,为进一步提升算法对拍摄角度、目标远近、录制过程抖动等因素的鲁棒性,分别将关键点坐标进行归一化、转化为相对坐标、标准化处理。关键点坐标的大小范围是相对于视频分辨率的,将坐标值 (x,y) 分别除以视频分辨率 $(v_{\mathrm{w}}, v_{\mathrm{h}})$ 归一化到(0,1)范围,减小了不同视频分辨率以及不同肢体关键点的数值差异。然后选取脖子部位的关键点 (x_0, y_0) 作为原点,对其他坐标进行变换:

$$(\bar{x}_i, \bar{y}_i)=(x_i, y_i)-(x_0, y_0) \tag{8}$$

式中:(x_i, y_i) 为人体关键点的坐标;$(\bar{x}_i, \bar{y}_i)$ 即为转换后的相对坐标。分类网络训练阶段对36个特征进行标准化处理,以降低不同样本的差异性。假设 p 代表同一时间步中样本的任一特征,即 $p \in \{\bar{x}_1, \bar{y}_1, \bar{x}_2, \bar{y}_2, \cdots, \bar{x}_{18}, \bar{y}_{18}\}$,定义公式如下:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{9}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (p_i - \mu)^2 \tag{10}$$

$$\bar{p}_i = \frac{p_i - \mu}{\sqrt{\sigma^2}} \tag{11}$$

式中:μ 为 n 个样本的均值;σ^2 为标准差。每个样本的特征通过式(11)进行标准化,$\bar{p}_i$ 则为转换后的鲁棒性特征。

2.2.3　分类网络设计

动作的描述可以由具备时序关系的一系列人体关键点构成[17]。为了充分挖掘序列的关系,设计了双层叠加的长短时记忆网络。本文设计的动作分类网络如图 5 所示,输入为滑窗提取并经过预处理得到的长度为 n 、特征维度为 36 的骨架关键点序列。两个 LSTM 网络单元的神经元个数为 32,时间步长等于滑动窗口的大小 n ,第一层隐藏层每个时间步的输出状态传递给第二层隐藏层,取最后一个时间步的输出向量传递给神经元个数为 64 的全连接层,最后通过 Softmax 分类器进行动作类型的识别。Softmax 分类器通过 Softmax 激活函数将多个神经元的输出映射到(0,1)之间,即各个类别的数值转化为概率,概率最大的类别即判定为分类结果。此外人体的大部分时间为静止坐位、慢走、站立等正常活动,而将正常活动误识别为康复动作对于患者的康复过程是不可靠的。分类网络将多种正常活动视为一类动作类型,同时进行康复动作以及正常活动的识别,提升了对康复动作识别的稳定性。

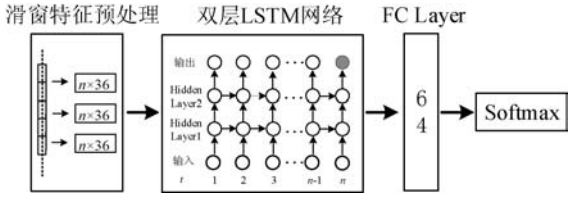

图 5　人体动作分类网络

网络参数的更新采用经过裁剪的康复动作视频段进行训练,即每个视频只包含特定的一个动作。由于视频时长不同而导致样本的时间步不同,通过 0 值填充转化为时间步一致的样本,训练时跳过特征值全为 0 的时间步。首先提取视频中人体的关键点序列,经过特征预处理转化为鲁棒性特征后输入到所设计的分类网络中,通过前向传播与反向传播完成参数的更新。结合 L2 正则化与 Dropout 来防止过拟合,训练完成后保存在测试集取得最优效果的参数。在线动作识别算法的分类网络加载训练好的参数,经过前向传播并通过 Softmax 分类器得出概率最大的动作类型,实时输出目标人体正在发生的动作类别与概率值。

3　实　验

3.1　实验平台

本实验的硬件环境如下:CPU 为酷睿 i7-8750,2.20 GHz,8 GB 内存;GPU 为 GTX 1060,6 GB 显存;监控摄像头的分辨率为 1 920 ×1 080。搭建的深度学习模型基于 TensorFlow 框架,采用 GPU 加速处理过程。

3.2　数据集

为了客观评价算法的准确性以及在康复场景的可用性,本文选取一个公开数据集 KTH[18],并参考文献[19]规范的脑卒中患者康复动作采集了一组康复动作数据集,各数据集如图 6 所示。

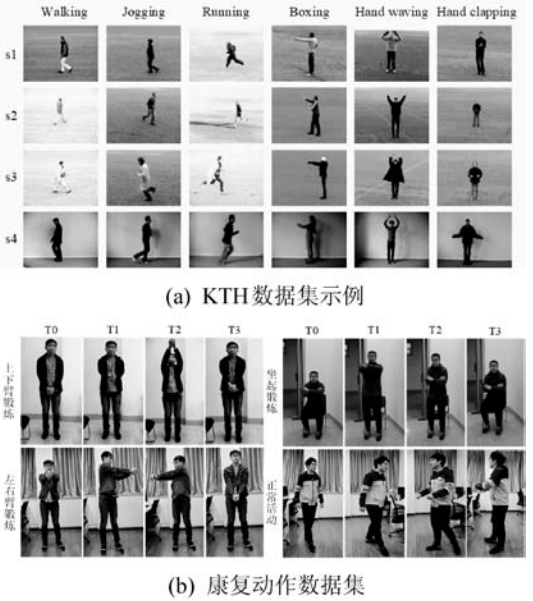

图 6　KTH 与康复动作数据集样本示例

1) KTH 是计算机视觉一个具有标志性的数据集,包含了 4 类场景下 25 个志愿者的 6 种行为:慢跑(Jogging)、步行(Walking)、跑步(Running)、拳击(Boxing)、挥手(Hand waving)和拍手(Hand clappong)。该数据集共有 599 个视频,每秒 25 帧,分辨率为 160 × 120,具有人体尺度和光照的变化,背景较为简单。按照与文献[18]一致的划分方式采用 18 个志愿者的视频作为训练集,其他 9 个志愿者为测试集。

2) 康复动作数据集由 5 位实验人员在 3 种不同的环境下采集而成,包含 4 类行为共 964 个视频段,具有光照、人体尺度、背景、拍摄角度的变化。行为类型分为三种康复动作以及一类正常活动动作,其中康复动作为上下臂锻炼、左右臂锻炼和坐起锻炼,正常活动包括静止站立、静止坐位、慢走、伸展等,具体动作说明见表 1。每个视频的分辨率为 1 280 ×720 或 1 920 ×

1 080,帧速率为 15 帧每秒,视频段持续时间为 8 ~ 15 s 之间。

表 1 脑卒中康复动作描述

动作类型	动作说明	目的
上下臂锻炼	双手交叉扣手,双臂伸直自腹部向头顶运动	肩关节前屈
左右臂锻炼	双手交叉扣手,双臂伸直后向左右运动	肩关节内收、外展
坐起锻炼	双臂交叉,由患侧负重站起	坐站训练

3.3 训练策略

两个数据集都为经过裁剪的短视频,数据集的基本训练流程如下:

1) 视频段中每帧提取的 36 个骨架关键点特征作为一个时间步,小于选定步长的样本通过补 0 的方式进行填充。

2) 对每帧提取的 36 个关键点特征进行预处理,将原始特征转化为鲁棒性特征。

3) 通过正态分布的方式生成随机值来初始化分类网络的权重参数,预处理后的样本分批量(batch-size)输入到分类网络,进行前向传播得到损失值,采用 Adam 优化器来最小化损失函数,学习率设置为 0.001。

KTH 数据集的样本较少,采用数据增广的方法将训练集扩充一倍,对视频进行左右对称变换并将其比例转换为 5∶4。将康复动作数据集按照 7∶3 的比例随机分为训练集与测试集,同时保证测试集中每种动作的样本比例均衡。另外,为了增大帧间动作差异同时提升运行效率,每间隔 3 个图像帧进行处理。训练时批量设置为 32,一共分 500 个 Epoch 运行,模型在 KTH 数据集和康复动作数据集上分别训练迭代9 000、6 000 次后逐渐收敛。

3.4 实验结果

3.4.1 不同模型设置对精度的影响

动作识别数据集的对比分析通常采用准确率作为评价标准,为分析不同的模型设置对识别精度的影响,实验分别从 LSTM 单元隐藏层节点个数、时间步长、特征预处理三个方面对 KTH 与康复动作数据集进行分析。实验分别将 LSTM 隐藏层节点个数设置为 16、32、64、128,时间步长统一设置为 50,实验结果如表 2 所示。当隐藏层节点个数依次增加时,KTH 与康复动作数据集的识别准确率分别提高至 94.9%、100%。依据实验结果选取最佳的隐藏层节点数量,在 KTH 数据集下的隐藏层节点个数设置为 64,康复动作数据集中设置为 32。

表 2 隐藏层节点个数对精度的影响 %

隐藏层节点个数	KTH	康复动作数据集
16	92.59	99.65
32	93.05	100
64	94.90	100
128	94.90	100

选取合理的时间步长对于识别精度是至关重要的,过短的时间步不能够充分表达一个动作,而过长的时间步则导致运算速度慢,冗余的信息也会干扰识别过程。实验分别将时间步长设置为 10、20、40、60、80,KTH 数据集中隐藏层节点个数设置为 64,康复动作数据集中隐藏层节点设置为 32,实验结果如表 3 所示。通过识别精度在每个数据集选取合理的时间步长,时间步长在 KTH 与康复动作数据集中分别大于 60、40 后模型的精度不再提高,即两个数据集分别通过提取的前 60 和 40 帧就能够达到最好的识别效果。

表 3 时间步长对精度的影响 %

时间步长	KTH	康复动作数据集
10	91.20	97.24
20	93.05	97.93
40	94.90	100
60	95.37	100
80	92.59	100

本文将姿态估计算法获取的骨架关键点进行了预处理,将其转换为鲁棒性特征,在对比分析加上预处理后的识别效果。KTH 数据集上时间步为 60,隐藏层节点个数为 64。康复动作数据集上时间步为 40,隐藏层节点个数为 32,对比分析结果如表 4 所示。相比于原始特征输入到分类模型,经过预处理后的鲁棒性特征在 KTH 与康复动作数据集的识别准确率分别提高了 2.78 和 1.04 个百分点。

表 4 特征预处理对识别精度的影响 %

特征类型	KTH	康复动作数据集
原始特征	92.59	98.96
鲁棒性特征	95.37	100

3.4.2 不同算法的识别精度对比

为了客观展示算法的性能,表 5 展示了与其他文献中的算法在 KTH 数据集上的对比结果。文献[7]是

在传统卷积神经网络的基础上增加了对时间维度的卷积,是动作识别领域的典型模型。文献[20]采用树状层次结构的深度网络提取视频的时空特征,结合 KNN 分类器进行动作识别。文献[21]提出融合兴趣点表现特征的增强单词包并通过 SVM 分类器实现动作识别。本文设计的动作识别算法均高于以上三种方法,识别准确率达 95.37%。由于 KTH 数据集分辨率较低,提取的骨架关键点存在较多丢失,同时算法在保证一定准确率的基础上提升了处理速度,识别精度略低于文献[22]方法。文献[22]首先采用 YOLO 算法[23]检测目标框,通过 CNN 提取目标框的图像特征并由 LSTM 网络进行分类,相比于本文算法,该特征提取模型更加复杂,计算量也更大。

表 5　KTH 数据集方法对比

方法	准确率/%
3D CNN[7]	90.20
Multi – level + KNN[20]	91.99
Improved BOW + SVM[21]	93.60
LC-YOLO[22]	96.63
本文方法	95.37

对于康复动作数据集,采用姿态估计算法提取骨架关键点并应用相同的特征预处理方式,分别与隐马尔可夫模型(HMM)、全连接循环神经网络(SimpleRNN)和门控循环单元(GRU)三种优秀的时序关系模型进行对比。对比分析结果如表 6 所示,相对于传统机器学习算法 HMM,循环神经网络对时序关系的提取能力更强,本文算法也取得了最好的识别结果。

表 6　康复动作数据集方法对比

方法	准确率/%
HMM	88.96
SimpleRNN	91.72
GRU	98.27
本文方法	100

3.4.3　在线康复动作识别

相比离线的数据集,在线动作识别更具挑战性。为测试康复训练场景下动作识别的效果,通过分辨率为 1 920 × 1 080 的单目摄像头捕获连续的视频流,采用在康复动作数据集中训练好的参数初始化分类网络。考虑到康复活动发生过程较慢,间隔 3 帧采样图像,采用时间步长为 80、隐藏层节点个数为 32 的分类网络参数。实时获取系统时间作为参考,在线动作识别效果如图 7 所示,图 7(a)、图 7(b)演示的动作分别为上下臂锻炼、左右臂锻炼;图 7(c)、图 7(d)演示的动作分别为正常活动、坐起锻炼,其中加入了无关人员的干扰,并且目标人体的位置产生了移动。算法处理能力达 18 帧每秒,能够持续捕捉并判断监控流中目标人体的康复动作,实时输出目标位置、动作类型以及动作概率,对于活动位置、其他人员干扰具有较强的适应能力。

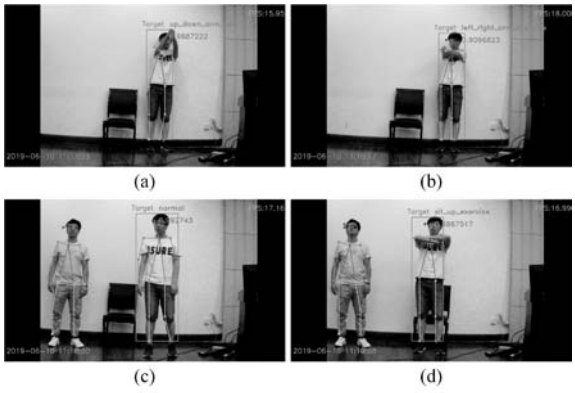

图 7　在线动作识别结果

为了客观展示算法于在线动作识别方式下对康复动作的识别准确率,实验人员连续做左右臂锻炼、上下臂锻炼、坐起锻炼各 50 次,并以站立、慢走等正常活动作为间隔动作。实验结果如表 7 所示,算法平均识别率达 93%,且不易将正常活动误识别为康复动作,在线场景下能够有效进行康复动作识别。然而相对于在康复动作数据集中的表现,在线场景下模型对三种康复动作的平均识别率仅 90.66%,原因在于实时环境下不同目标的动作行为存在较大的不确定性,需要更加充分的数据集训练分类网络来进一步达到更好的识别效果。

表 7　连续动作识别结果

动作类型	正确识别次数	错误识别次数	准确率/%
左右臂锻炼	43	7	86
上下臂锻炼	49	1	98
坐起锻炼	44	6	88
正常活动	150	0	100

4　结　语

本文提出了一种基于单目视觉的在线动作识别算法,结合姿态估计 OpenPose 与最近邻匹配算法对视频流中的目标人体生成动作序列,采用滑动窗口选取原始关键点特征并转换为鲁棒性特征后输入到双层

LSTM 网络进行动作分类。利用 OpenPose 获取骨架关键点的实时性对监控流中的目标人体进行快速检测，同时基于基准坐标结合最近邻匹配算法跟踪目标人体，避免了视频流中大量无关区域以及其他人体的噪声干扰。结合 LSTM 对长时间序列的处理能力，能够对视频流中目标人体的行为做出准确的识别。通过在公开数据集 KTH 和康复动作数据集中实验，KTH 数据集的平均识别率达 95.37%，康复动作数据集中的识别率达到了 100%。在线动作识别下的康复动作识别率达 90.66%，证明了该方法的有效性，探索了基于计算机视觉的动作识别方法在康复训练领域的应用。

由于康复动作数据集的样本规模较小，算法在连续视频流中进行在线动作识别下的识别率与其在数据集中的表现存在较大差距。未来研究将会采集更丰富的数据集，加入更多的康复动作类型，同时优化训练策略与分类网络，进一步提升对真实场景的适应能力，帮助患者在居家场景下更好地完成康复训练计划。

参 考 文 献

[1] Zhang B W, Wang L M, Wang Z, et al. Real-time action recognition with deeply-transferred motion vector CNNs[J]. IEEE Transactions on Image Processing, 2018, 27(5): 2326-2339.

[2] 朱翠平，吴美华，徐晓芳，等. 家庭康复护理对农村脑卒中偏瘫病人肢体运动功能的影响[J]. 护理研究，2017，31(11)：1365-1367.

[3] 盛晗，邵圣文，王惠琴，等. 脑卒中患者康复锻炼依从性动态变化的研究[J]. 中华护理杂志，2016，51(6)：712-715.

[4] Bisio I, Delfino A, Lavagetto F, et al. Enabling IoT for in-home rehabilitation: Accelerometer signals classification methods for activity and movement recognition[J]. IEEE Internet of Things Journal, 2016, 4(1): 135-146.

[5] 马高远，林明星，吴筱坚，等. 基于人体动作反馈的上肢康复机器人主动感知系统[J]. 机器人，2018，40(4)：491-499.

[6] Wang H, Schmid C. Action recognition with improved trajectories[C]//2013 IEEE International Conference on Computer Vision. IEEE, 2013: 3551-3558.

[7] Ji S W, 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.

[8] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. ACM, 2014: 568-576.

[9] Donahue J, Hendricks L A, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 677-691.

[10] Xu H J, Das A, Saenko K. R-C3D: Region convolutional 3D network for temporal activity detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 5794-5803.

[11] 罗会兰，王婵娟，卢飞. 视频行为识别综述[J]. 通信学报，2018，39(6)：169-180.

[12] 邵阳，战荫伟，许碧雅. 余弦 DTW 在上肢康复训练中的应用[J]. 计算机工程与设计，2018，39(1)：249-254.

[13] 唐心宇，宋爱国. 人体姿态估计及在康复训练情景交互中的应用[J]. 仪器仪表学报，2018，39(11)：195-203.

[14] Gama A E F D, Chaves T D M, Fallavollita P, et al. Rehabilitation motion recognition based on the international biomechanical standards[J]. Expert Systems with Applications, 2019, 116: 396-409.

[15] Li Y H, Lan C L, Xing J L, et al. Online human action detection using joint classification-regression recurrent neural networks[C]//2016 14th European Conference on Computer Vision. Springer, 2016: 203-220.

[16] Cao Z, Hidalgo G, Simon T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172-186.

[17] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//32nd AAAI Conference on Artificial Intelligence, 2018.

[18] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach[C]//Proceedings of the 17th International Conference on Pattern Recognition. IEEE, 2004: 32-36.

[19] 许梦雅，杨伟民. 家庭医疗体操在缺血性脑卒中社区康复中的应用[J]. 中国老年学，2010，30(17)：2437-2438.

[20] Charalampous K, Gasteratos A. On-line deep learning method for action recognition[J]. Pattern Analysis and Applications, 2016, 19(2): 337-354.

[21] 张良，鲁梦梦，姜华，等. 局部分布信息增强的视觉单词描述与动作识别[J]. 电子与信息学报，2016，38(3)：549-556.

[22] 马钰锡，谭励，董旭，等. 面向智能监控的动作识别[J]. 中国图象图形学报，2019，24(2)：128-136.

[23] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 6517-6525.