计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300042

基于骨骼关键点检测的多人行为识别

李梦荷　许宏吉　石磊鑫　赵文杰　李　娟
山东大学信息科学与工程学院　山东 青岛 266237
(limenghe0309@163.com)

摘　要　人体行为识别(Human Activity Recognition,HAR)技术是计算机视觉领域的研究热点,目前多人 HAR 的研究仍存在很多技术难点。针对多人 HAR 中人数判断不准确、特征提取难度大导致行为识别准确率低的问题,提出了一种基于骨骼关键点检测的多人行为识别系统。该系统将骨骼点提取与动作识别相结合,首先对原始视频进行图像帧提取,然后通过 OpenPose 算法得到人体骨骼关键点数据来对人体进行检测并标注,最后根据骨骼点的特点提取人体姿态特征。同时,为准确描述特征之间的关系,提出了一种基于帧窗口矩阵的特征描述方法,该方法将支持向量机(Support Vector Machine,SVM)作为分类器以完成多人行为识别。选择 UT-Interaction 和 HMDB51 这两个公开的数据集中的 10 类日常典型行为作为测试对象,实验结果表明,所提方法可以有效提取图像中的多人骨骼关键点信息,且其对 10 类日常典型行为的平均识别准确率达 86.25%,优于对比的其他已有方法。

关键词:OpenPose 算法;骨骼关键点提取;姿态特征提取;SVM 分类器
中图法分类号　TP391

Multi-person Activity Recognition Based on Bone Keypoints Detection

LI Meng-he,XU Hong-ji,SHI Lei-xin,ZHAO Wen-jie and LI Juan
School of Information Science and Engineering,Shandong University,Qingdao,Shandong 266237,China

Abstract　Human activity recognition(HAR) technology is a research hotspot in the field of computer vision,but there are still many technical difficulties in the research of multi-person HAR. The problem of the inaccurate judgment of the number of people and the difficulty of feature extraction in multi-person activity recognition may lead to the low accuracy. A multi-person activity recognition system based on bone keypoints detection is proposed in this paper,which combines the extraction of bone points with the action recognition. Firstly,the image frame is extracted from the original video. Secondly,the OpenPose algorithm is used to obtain keypoints data of the human skeleton to detect the number of people in the image and mark activity information. At last,human posture features are extracted according to characteristics of skeleton points. Meanwhile,in order to accurately describe the relationship between posture features,a feature description method based on frame window matrix is proposed. Finally,a support vector machine(SVM) is used as a classifier to complete multi-person activity recognition. 10 types of daily typical activities from UT-Interaction and HMDB51 datasets are taken as test objects,and experimental results prove that the proposed method can effectively extract keypoints of multiple human bones in the image. Its average recognition accuracy of 10 activities is 86.25%,which is higher than other compared methods.

Keywords　OpenPose algorithm,Skeleton keypoints extraction,Posture feature extraction,SVM Classifier

1　引言

近年来,随着计算机视觉的发展,人体行为识别技术迅速发展并得到了广泛应用[1]。目前 HAR 的研究主要集中在单人的行为识别方面,多人行为识别技术的算法相对较少,但在现实生活中多人行为或群体行为所占比重却很大[2]。多人 HAR 技术可被广泛应用于智能家居、智慧医疗、异常行为监控、体育竞技和智慧课堂等领域[3-5]。因此,多人 HAR 技术的研究有着重要的理论意义和实际应用价值。

多人 HAR 是指两人或两人以上的行为识别。在以往对多人 HAR 的研究中,针对特征层面,Sun 等[6]提出了改进的运动历史图和基于能量块的二维高斯滤波器的方法对人体的动作信息进行特征提取;Tu 等[7]提出了一种新的局部聚集描述符的时空向量方法,该方法可以自适应地提取视频里的深

到稿日期:2020-03-09　返修日期:2020-08-02　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:国家重点研发计划子课题(2018YFC0831001);国家自然科学基金(61771292);山东省教育科学"十三五"规划课题(YZ2019070)
This work was supported by the National Key Research and Development Program of China(2018YFC0831001),National Natural Science Foundation of China(61771292) and the 13th Five-Year Plan on Education Science of Shandong Province(YZ2019070).
通信作者:许宏吉(hongjixu@sdu.edu.cn)

层特征并进行调整，再利用相似度权重来汇总信息量大的特征图。针对系统框架层面，Bagautdinov 等[8]提出了一种端到端的框架，该框架可以联合检测多个个体，然后提取各个个体的密集特征，通过匹配递归神经网络来进行分析处理。

目前，研究人员大多基于目标检测来进行多人行为识别的建模和研究[9-13]，其难点在于由于无法准确判断图像中的人数，不能提取出有效的特征，最终导致行为识别的准确率不高。针对这一问题，本文提出了一种基于骨骼关键点检测的多人行为识别方法，该方法通过姿态估计算法来提取骨骼关键点，并与动作识别相结合，进行图像中人数的检测、人体姿态特征的提取，从而有效提高了行为识别的准确率。

2 系统框架

2.1 系统介绍

基于骨骼关键点检测的多人 HAR 系统的实现框架如图 1 所示，其主要包括图像预处理模块、骨骼关键点检测模块、特征提取模块和行为识别模块这 4 个部分。其中，骨骼关键点检测模块包含外部特征提取和骨骼点提取（多人检测）；特征提取模块包含多人姿态特征提取和特征点描述。该系统首先通过 OpenPose 算法进行多人骨骼点的检测，然后对骨骼点数据进行特征提取与特征描述，最后与支持向量机分类算法相结合以实现多人行为识别。

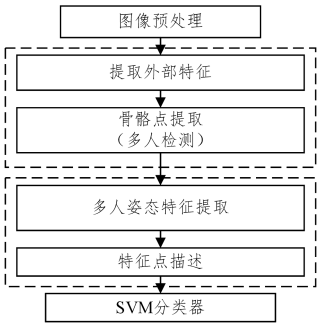

图 1 基于骨骼关键点检测的多人 HAR 系统框架

Fig. 1 Framework of multi-person HAR system based on bone keypoint detection

2.2 系统流程

基于骨骼关键点检测的多人 HAR 系统处理流程如图 2 所示，其主要处理步骤如下。

Step 1 图像预处理

首先将原始视频帧转换为图像帧，然后通过滤波器对图像进行降噪处理来消除图像中无关信息对后续结果的影响。

Step 2 训练集和测试集的分配及动作标注

将 80％的数据集作为训练集，20％的数据集作为测试集，并对每类动作进行类别标注。

Step 3 提取骨骼关键点

首先通过视觉几何组神经网络的前 19 层（Visual Geometry Group Network 19，VGG-19）来提取训练集数据的外部特征，然后在此基础上，针对每类动作的每帧图像提取主要骨骼关键点，并进行格式转换，最后形成骨骼关键点的文本文件。

Step 4 骨骼关键点的优化与特征提取

首先对骨骼关键点进行优化处理，主要包括主关键点的确定和部分关键点缺失的补充这两部分。然后结合骨骼点的空间特点，提取人体姿态特征。最后利用一种基于帧窗口矩阵的特征点描述方法来确定一个较完整的输入特征矩阵。

Step 5 训练与分类

调用 SVM 分类器进行多人行为的识别。

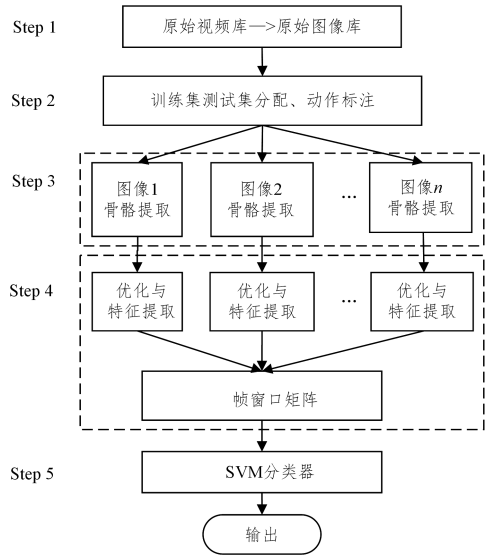

图 2 基于骨骼关键点检测的多人 HAR 系统处理流程

Fig. 2 Flowchart of multi-person HAR system based on bone keypoint detection

3 模型与方法

3.1 骨骼关键点的检测

骨骼关键点检测模块主要检测人体重要的关键点，如关节点、五官节点等。OpenPose 算法最初是以 Caffe 为框架的开源人体检测项目，其采用自下而上的方法，能够对面部关键点、人体关键点和手部关键点进行检测，适用于单人和多人的检测，具有很强的鲁棒性。

OpenPose 算法的网络结构如图 3 所示[14-15]。

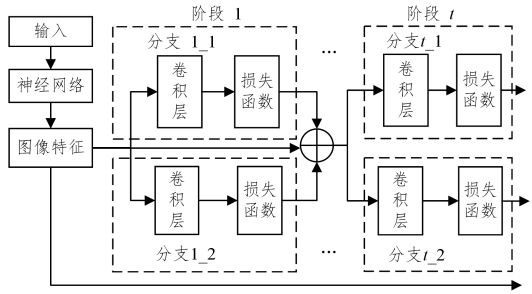

图 3 OpenPose 网络结构[15]

Fig. 3 Structure of OpenPose network[15]

具体处理流程如下。首先通过 VGG-19 神经网络进行训练，获得图像的外部特征。然后进入预测阶段，每个预测阶段被分为两个分支进行迭代训练，分支 1_1 通过多个卷积层来预测人体各个骨骼关键点的位置；分支 1_2 通过多个卷积层来预测各个骨骼关键点之间的连接信息。每个预测阶段都会计算一次损失函数，之后将 VGG-19 提取到的图像特征、骨骼

点位置信息、骨骼点之间的连接信息融合，并送入下一个阶段再进行训练。如此反复，最后经过 $t(t\geqslant 2)$ 个阶段，得到图像中人体的骨骼关键点。

分支 t_1、分支 t_2 和总的损失函数的计算如式(1)－式(3)所示：

$$f_S^t=\sum_{j=1}^{J}\sum_{P}\boldsymbol{W}(p)\cdot\left\|S_j^t(p)-S_j^*(p)\right\|_2^2 \tag{1}$$

$$f_L^t=\sum_{i=1}^{C}\sum_{p}\boldsymbol{W}(p)\cdot\left\|L_c^t(p)-L_c^*(p)\right\|_2^2 \tag{2}$$

$$f=\sum_{t=1}^{T}(f_S^t+f_L^t) \tag{3}$$

其中，f_S^t 是 t 阶段分支 t_1 的损失函数值，表示 t_1 输出的预测值 S_j^t 与标注值 S_j^* 之间的 L_2 范数；j 为骨骼点序号；t 表示预测阶段序号。$\boldsymbol{W}(p)$ 为二值矩阵，当其取 0 时表示某个关键点标注缺失，损失函数不计算该点；当其取 1 时表示关键点标注正常。f_L^t 是 t 阶段分支 t_2 的损失函数值，表示 t_2 输出的预测值 L_c^t 与标注值 L_c^* 之间的 $L2$ 范数；c 表示关节连接号。f 表示训练阶段总的损失函数值，是 f_S^t 和 f_L^t 的总和。

经过 OpenPose 算法检测后，得到人体的骨骼关键点，如图 4 所示，对应的 18 个骨骼点序号如表 1 所列。

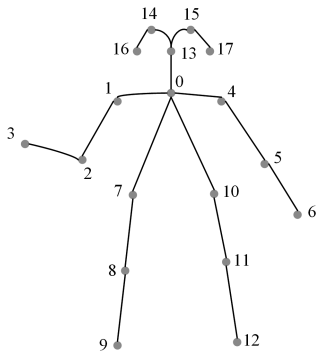

图 4 人体骨骼关键点示意图

Fig. 4 Diagram of keypoints of human skeleton

表 1 18 个人体骨骼关键点的定义

Table 1 Definition of 18 bone keypoints of human skeleton

序号	骨骼点编号	位置
0	No. 0 Neck	脖子
1	No. 1 L-Shoulder	左肩
2	No. 2 L-Elbow	左半臂
3	No. 3 L-Wrist	左手腕
4	No. 4 R-Shoulder	右肩
5	No. 5 R-Elbow	右半臂
6	No. 6 R-Wrist	右手腕
7	No. 7 L-Hip	左胯骨
8	No. 8 L-Knee	左膝盖
9	No. 9 L-Ankle	左脚腕
10	No. 10R-Hip	右胯骨
11	No. 11 R-Knee	右膝盖
12	No. 12 R-Ankle	右脚腕
13	No. 13 Nose	鼻子
14	No. 14 L-Eye	左眼
15	No. 15 R-Eye	右眼
16	No. 16 L-Ear	左耳
17	No. 17 R-Ear	右耳

3.2 骨骼点数据的改进

通过 OpenPose 算法实现从图像格式(.jpg)到骨骼数据格式(.txt)的转换。转换后得到人体的 18 个骨骼关键点数据，每个关键点由 x 轴坐标和 y 轴坐标组成(以图像的左下角为坐标原点)，一维向量的格式为 $[x_0,x_1,\cdots,x_i,\cdots,x_n]$，$0\leqslant i\leqslant 35$。其中，$x_0,x_1$ 分别表示 No. 0 Neck(脖子)的 x 轴和 y 轴数据，x_2,x_3 分别表示 No. 1 L-Shoulder(左肩)的 x 轴和 y 轴数据，以此类推。

3.2.1 骨骼点数据的优化

由于同一个动作的不同帧之间可能会出现漏检的情况，需要先对骨骼关键点数据进行优化，优化过程如下。

1)确定主关键点。由图 4 可知，因为脖子关键点连接所有的关节，为核心关键点，所以将脖子关键点定为主关键点。在一套动作中，若某个人所有的帧都未检测到主关键点，即缺失 No. 0 Neck(脖子)的数据，则剔除该套动作，并在剔除后，根据检测到的 No. 0 Neck 数重新确认图像中的人数。

2)对部分关键点缺失的补充。若某一帧图像缺失一个或多个关键点，则根据前后两帧图像关键点位置的均值来进行填充，具体如式(4)和式(5)所示。

$$x_i=\frac{x_{i-1}+x_{i+1}}{2} \tag{4}$$

$$y_i=\frac{y_{i-1}+y_{i+1}}{2} \tag{5}$$

其中，i 表示关键点序号，$2\leqslant i\leqslant 35$。

若第一张和最后一张图像缺失关键点，则直接将其删除。若由于遮挡原因无法检测到腿部、胯部的关键点，则进行补 0 处理。

3.2.2 特征提取

特征提取在多人 HAR 中占有重要位置，特征的有效性将直接影响后续的识别结果。因为在进行骨骼点提取之前已经通过 VGG-19 网络提取了图像的外部特征，所以在此基础上本节将根据骨骼关键点的空间特征，进行二次特征提取，即提取人体的姿态特征，其主要包含以下 3 个部分。

1)计算偏移量

偏移量表示各个关键点与主关键点(No. 0 Neck)的距离，如式(6)和式(7)所示：

$$X_{\text{offset}}=x_i-x_{\text{neck}} \tag{6}$$

$$Y_{\text{offset}}=y_i-y_{\text{neck}} \tag{7}$$

其中，$X_{\text{offset}},Y_{\text{offset}}$ 分别表示在 x 轴和 y 轴上的偏移量，(x_i,y_i) 表示骨骼关键点坐标。

2)提取骨骼关键点的 3 个特征

骨骼关键点数据的 3 个特征包括：人体运动速度 V_{body}、骨骼点偏移角运动速度 V_{angle} 和骨骼点运动速度 V_{joint}，具体计算式如式(8)－式(10)所示：

$$V_{\text{body}}^i=\frac{x_{\text{neck}}^{t_i}-x_{\text{neck}}^{t_{i-1}}}{t_i-t_{i-1}} \tag{8}$$

其中，i 表示当前点的坐标序号。

$$V_{\text{angle}}^i=\frac{\arctan\left(\frac{\Delta Y_{\text{offset}}}{\Delta X_{\text{offset}}}\right)}{t_i-t_{i-1}} \tag{9}$$

其中，$\Delta Y_{\text{offset}}=Y_{\text{offset}}^{t_i}-Y_{\text{offset}}^{t_{i-1}}$，$\Delta X_{\text{offset}}=X_{\text{offset}}^{t_i}-X_{\text{offset}}^{t_{i-1}}$。

$$V_{\text{joint}}^i=\frac{x_i^{t_i}-x_{i-1}^{t_{i-1}}}{t_i-t_{i-1}} \tag{10}$$

3)基于帧窗口参数的特征描述方法

提取特征后，组成一维输入特征向量。动作识别需要对完整的动作进行处理，而特征向量仅能表示人体此刻的动作，不能确定一连贯的行为信息之间的关系，即每条特征数据相对独立。为了确定一个连惯的动作之间有多少条数据，即确定需要多少帧图像，本文提出了一种基于帧窗口矩阵的特征描述方法来表示一个较完整的动作，帧窗口（***frame_window***）类似于滑动窗口，表示一个较完整的动作（N 帧），帧窗口矩阵的一行表示一个骨骼关键点提取特征后的数据，一列表示该动作有多少帧。

帧窗口 $\boldsymbol{F}_j$ 矩阵的表示如式(11)所示：

$$\boldsymbol{F}_j=\begin{bmatrix}\boldsymbol{Feature}_0^j\\ \boldsymbol{Feature}_1^j\\ \cdots\\ \boldsymbol{Feature}_i^j\\ \cdots\\ \boldsymbol{Feature}_N^j\end{bmatrix} \tag{11}$$

其中，N 表示帧窗口矩阵的行数，即表示一个较完整动作所包含的帧数；j 表示动作的矩阵数；$\boldsymbol{Feature}_N^j$ 表示提取的骨骼关键点的一维特征向量，$\boldsymbol{Feature}_N^j$ 的公式如式(12)所示：

$$\boldsymbol{Feature}_N^j=V_{bodyN}^j,V_{angleN}^j,\cdots,V_{jointN}^j,\cdots,V_{jointN}^j \tag{12}$$

其中，各个特征数据根据提取的人体运动速度 V_{body}、骨骼点偏移角速度 V_{angle}、骨骼点运动速度 V_{joint} 的顺序进行排列，且每个关键点按表 1 列出的 No. 0－No. 17 的编号从小到大组成 $\boldsymbol{Feature}_N^j$ 向量。

3.3 行为分类

将特征矩阵输入 SVM 分类器进行分类识别，以验证所提方法的有效性。

4 实验结果与分析

4.1 实验环境

本文的仿真实验环境如表 2 所列。

表 2 实验环境介绍

Table 2 Introduction of experimental environment

	名称	介绍
硬件环境	处理器	Intel Xeon CPU E5-2678 v3 @2.50GHz
	显卡	NVIDIA GeForce GTX 1060
	内存	16.0GB
软件环境	操作系统	Win 10 企业版
	开发语言	Python
	开发环境	Anaconda3＋Python3.6＋CUDA9.0＋cudnn＋Tensorflow-gpu 1.8
	第三方库	OpenCV 4.1＋Numpy 1.15＋Sklearn 0.21＋Matplotlib 2.1＋Shaply 1.6 等

4.2 公开数据集

本文选择 UT-Interaction 和 HMDB51 这两个公开数据集作为测试对象。

UT-Interaction 数据集[16]于 2010 年发布，包含拥抱、踢、指、殴打、推、握手这 6 类人体交互行为，视频分辨率为 720×480 像素；HMDB51 数据集[17]于 2011 年发布，包含拍手、走、运球、打高尔夫等 51 类日常生活中常见的行为，视频分辨率为 320×240 像素，该数据集所有的视频均截取自电影片段，包含很多实际应用场景中的多类行为。

基于日常生活中的典型应用场景，本文选择 UT-Interaction 中的多人交互行为（拥抱、握手）、异常行为（踢、指、殴打、推），以及 HMDB51 数据集中的 10 类具有代表性的日常行为（拍手、走、运球、打高尔夫等）作为测试对象。本文先用 Python＋OpenCV 进行预处理，实现视频到图像的动作帧截取，截取速率为 15 帧/秒，并通过 OpenCV 来显示图像，以便下一步进行行为识别处理。

4.3 骨骼关键点检测结果与分析

选择 HMDB51 数据集、UT-Interaction 数据集中的多个图像作为仿真实验的输入，对手部关节、面部关节及全身关节进行检测，实验结果如图 5 所示。

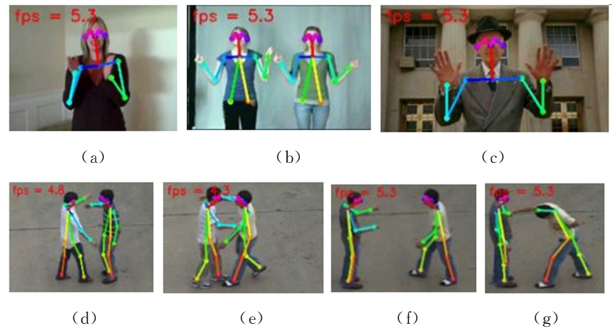

图 5 骨骼点检测结果

Fig. 5 Results of bone point detection

其中，图 5(a)－图 5(c)来自 HMDB51 数据集，分别代表单人鼓掌、两人鼓掌、单人挥手行为；图 5－图 5(g)来自 UT-Interaction 数据集，图 5(d)、图 5(e)代表拥抱行为，图 5(f)、图 5(g)代表打架行为。

虽然图 5 中的图像来自不同的公开数据集，图像的分辨率不同，但是基于 OpenPose 的骨骼关键点检测方法依然可以有效地提取人体的骨骼关键点，从而对目标图像中的人数进行确定，且检测效果良好。

4.4 对比实验与分析

4.4.1 帧窗口参数 f 实验对比

经过多次实验，对帧数进行统计，在所有统计的完整动作中，最少的动作帧数是 18，因此为有效利用时间和系统内存空间，设置帧窗口的参数值为[10,18]，分别在 10 类行为的训练集上和测试集上进行仿真实验，识别准确率如图 6 所示。

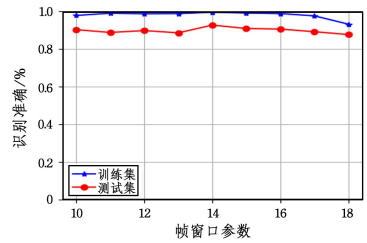

图 6 不同帧窗口参数值识别准确率的对比

Fig. 6 Comparison of recognition rate for different frame window parameter values

图 6 给出了当 10 类行为取不同帧窗口参数值时，其在训练集上和测试集上行为的平均识别准确率的变化曲线，曲线的横坐标表示帧窗口的参数值，曲线的纵坐标表示行为识别准

确率。由图6可知，当 $f=14$ 时，训练集获得最高的识别准确率，值为98.3%；测试集获得最高的识别准确率，值为92.5%。

在多人交互行为（拥抱、握手）、异常行为（踢、指、殴打、推）、日常行为（拍手、走、运球、打高尔夫）的数据上分别进行帧窗口参数的仿真实验对比，得到的识别准确率如表3－表5所列。

表3 多人交互行为训练集和测试集的准确率对比

Table 3 Comparison of accuracy on training sets and testing sets of multi-person interaction activities

帧窗口值	训练集		测试集	
	拥抱	握手	拥抱	握手
10	0.87	0.91	0.84	0.89
11	0.92	0.94	0.87	0.90
12	0.92	0.97	0.89	0.93
13	0.93	0.98	0.90	0.95
14	**0.94**	**0.99**	**0.92**	**0.97**
15	0.91	0.96	0.89	0.94
16	0.90	0.94	0.88	0.90
17	0.92	0.94	0.87	0.89
18	0.89	0.92	0.84	0.88

表4 异常行为训练集和测试集上的准确率对比

Table 4 Comparison of accuracy on training sets and testing sets of abnormal activities

帧窗口值	训练集				测试集			
	踢	指	殴打	推	踢	指	殴打	推
10	0.86	0.94	0.75	0.83	0.84	0.90	0.65	0.79
11	0.87	0.96	0.72	0.80	0.84	0.93	0.68	0.80
12	0.88	0.94	0.77	0.81	0.85	0.94	0.74	0.81
13	0.90	0.89	0.73	0.83	0.87	0.96	0.70	0.80
14	**0.92**	**0.98**	**0.79**	**0.86**	**0.90**	**0.98**	**0.78**	**0.84**
15	0.90	0.97	0.75	0.80	0.89	0.95	0.72	0.83
16	0.89	0.94	0.64	0.90	0.87	0.90	0.60	0.80
17	0.84	0.89	0.52	0.86	0.79	0.89	0.50	0.77
18	0.74	0.90	0.50	0.88	0.70	0.84	0.50	0.70

表5 日常行为训练集和测试集的准确率对比

Table 5 Comparison of accuracy on training sets and testing sets of daily activities

帧窗口值	训练集				测试集			
	拍手	走	运球	打高尔夫	拍手	走	运球	打高尔夫
10	0.79	0.91	0.77	0.76	0.84	0.89	0.68	0.73
11	0.81	0.94	0.74	0.80	0.84	0.90	0.70	0.75
12	0.84	0.96	0.79	0.80	0.85	0.92	0.73	0.79
13	0.83	0.97	0.83	0.82	0.87	0.89	0.77	0.80
14	**0.89**	**0.99**	**0.85**	**0.87**	**0.90**	**0.95**	**0.78**	**0.85**
15	0.87	0.93	0.85	0.85	0.89	0.94	0.74	0.83
16	0.84	0.95	0.81	0.80	0.87	0.90	0.70	0.80
17	0.82	0.91	0.72	0.83	0.79	0.87	0.69	0.75
18	0.70	0.90	0.70	0.79	0.70	0.86	0.64	0.73

由图6、表3－表5可以看到，各种行为在训练集上和测试集上的识别准确率的对比结果，最终本文选取 $f=14$ 作为帧窗口矩阵的参数。

4.4.2 典型方法的性能对比

为便于与已有算法进行性能对比，本文在HMDB51数据集上将所提出的基于骨骼关键点检测的多人行为识别方法与Carreira等[18]提出的Two-Stream Inflated 3D ConvNets（T-SI3D）方法、Choutas等[19]提出的Inflated 3D ConvNets PoTion（I3D PoTion）方法和Wang等[20]提出的Hallucinating IDT Descriptors and I3D Optical Flow Features方法进行对比，实验结果如表6所列。

表6 本文方法与其他典型方法的对比

Table 6 Comparison of proposed method with other typical methods

方法	平均识别准确率/%
Carreira等[18]	80.70
Choutas等[19]	80.90
Wang等[20]	82.37
本文方法	**86.25**

由表6可知，本文方法较其他方法在识别准确率上有一定提升，与Wang等[20]的方法相比，其识别准确率提高了3.88%。

结束语 本文结合姿态估计算法，通过OpenPose检测人体骨骼关键点，并提取骨骼点姿态特征，利用帧窗口矩阵和SVM分类器实现了多人行为的准确识别，对10类日常典型行为的平均识别准确率达到87.50%，在HMDB51数据集上的识别准确率达到86.25%。实验结果表明，所提基于骨骼关键点检测的多人行为识别系统可以有效解决图像中人数判定不准确、特征提取难度大的问题，进而提高了多人行为识别的准确率。

本文对未来工作的展望：

1）根据不同动作帧之间的关系，寻找更加合理高效的特征描述方法；

2）利用本文数据和方法对图神经网络模型（Graph Neural Networks，GNN）进行研究。

参 考 文 献

[1] LIU A A, SU Y T, JIA P P, et al. Multiple/Single-View Human Action Recognition via Part-Induced Multitask Structural Learning [J]. IEEE Transactions on Cybernetics, 2015, 45(6): 1194-1208.

[2] GONG W. Design and Implementation of Student Learning Behavior Recognition System Based on Skeleton Keypoint Detection [D]. Changchun: Jilin University, 2019.

[3] DAWAR N, KEHTARNAVAZ N. Action Detection and Recognition inContinuous Action Stream by Deep Learning-Based Sensing Fusion [J]. IEEE Sensors Journal, 2018, 18(23): 9660-9668.

[4] CHENG J, LIU H J, WANG F, et al. Silhouette Analysis for Human Action Recognition Based on Supervised Temporal T-SNE and Incremental Learning [J]. IEEE Transactions on Image Processing, 2015, 24(10): 3203-3217.

[5] LIU A A, XU N, NIE W Z, et al. Multi-Domain and Multi-Task Learning for Human Action Recognition[J]. IEEE Transactions on Image Processing, 2019, 28(2): 853-867.

[6] SUN J F, XU H J, ZHOU Y M, et al. Human Actions Recognition Using Improved MHI and 2-D Gabor Filter Based on Energy Blocks [C]// 2018 International Conference on Artificial Intelligence: Technologies and Applications (ICAITA2018). Chengdu: Atlantis Press, 2018: 1-4.

[7] TU Z G, LU H Y, ZHANG D J, et al. Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition [J]. IEEE

Transactions on Image Processing，2019，28(6)：2799-2812.

[8] BAGAUTDINOV T，ALAHI A，FLEURET F，et al. Social Scene Understanding：End-to-End Multi-Person Action Localization and Collective Activity Recognition [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017). Hawaii：IEEE，2017：3425-3434.

[9] ZHOU Q Q，ZHONG B N. Deep Alignment Network Based Multi-Person Tracking with Occlusion and Motion Reasoning [J]. IEEE Transactions on Multimedia. 2019，21(5)：1183-1194.

[10] LI M P，ZHOU Z M，et al. Multi-Person Pose Estimation Using Bounding Box Constraint and LSTM [J]. IEEE Transactions on Multimedia. 2019，21(10)：2653-2263.

[11] LIN L，WANG Y F，et al. Multi-Person Pose Estimation Using Aurous Convolution [J]. Electronics Letters. 2019，55(9)：533-535.

[12] CHEN X，YANG G K. Multi-Person Pose Estimation with LIMB Detection Heatmaps [C]// 2018 IEEE International Conference on Image Processing(ICIP 2018). Athens：IEEE，2018：4078-4082.

[13] ANDRILUKA M，ROTH S，SCHIELE B. Pictorial Structures Revisited：People Detection and Articulated Pose Estimation [C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2009). Miami，FL：IEEE，2009：1014-1021.

[14] CAO Z，SIMON T，WEI S E，et al. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017). Hawaii：IEEE，2017：1302-1310.

[15] CAO Z，HIDALGO G，SIMON T. OpenPose：Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields [C]// 2019 IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2019). Hawaii：IEEE，2019：1-14.

[16] RYOO M S，AGGARWAL J K. Spatio Temporal Relationship Match：Video Structure Comparison for Recognition of Complex Human Activities [C]// 2009 IEEE International Conference on Computer Vision(CVPR 2009). 2009：1593-1600.

[17] YAN S J，XIONG Y J，LIN D H. Spatial Temporal Graph Convolutional Networks for Skeleton Based Action Recognition [C]// 2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2018). Salt Lake City：IEEE，2018：7444-7452.

[18] CARREIRAL J，ZISSENRMAN A. Quo Vadis：Action Recognition? A New Model and the Kinetics Dataset [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017). Hawaii：IEEE，2017：4724-4733.

[19] CHOUTAS V，WEINZAEPFEL P，REVAUD J. Potion：Pose-Motion Representation for Action Recognition [C]// 2018 Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City：IEEE，2018：7024-7033.

[20] WANG L，KONIUSZ P，HUYNH D Q. Hallucinating IDT Descriptors and I3D Optical Feature for Action Recognition with CNNs [C]// 2019 IEEE International Conference on Computer Vision(ICCV 2019). Seoul：IEEE，2019：1-12.

LI Meng-he，born in 1994，postgraduate. Her main research interests include computer vision and artificial intelligence.

XU Hong-ji，born in 1976，Ph.D，associate professor. His main research interests include wireless communications，ubiquitous computing，intelligent perception，blind signal processing and artificial intelligence.