◆人工智能算法与应用◆

基于轻量化二维人体姿态估计的小样本动作识别算法

尹继尧[1**],周　琳[1],李　强[1],刘董经典[2]
(1. 深圳市城市公共安全技术研究院,广东深圳　518046;2. 中国矿业大学计算机科学与技术学院,江苏徐州　221116)

摘要:动作识别是近年来时序数据挖掘领域的研究热点,具有广泛的应用前景。但是现阶段基于深度学习的动作识别算法需要大量的标记训练数据集,存在泛化性差、实时性差、场景受限的问题。为解决这些问题,本研究设计一种基于轻量化二维人体姿态估计的小样本动作识别算法。该算法基于 YOLOv5 算法构建轻量化的人体检测器 HYOLOv5。基于轻量化二维姿态估计模型 Lite-HRNet 设计人体姿态特征描述算子,有效地去除背景对人体动作特征的干扰。为有效度量时序人体姿态特征描述算子间的相似度,本研究提出基于动态时间规整的人体姿态特征距离度量,并在此基础上设计基于类别中心选择的动作模板匹配算法。该算法通过少量的动作视频构建动作特征模板库,利用动作模板匹配算法可实现多类动作视频的精准识别。为验证算法,本研究在 COCO 2017 的 Humans 数据集上对 HYOLOv5 进行测试,人体检测识别精度 mAP@0.5∶0.95 可达 50.7%。基于 10 种动作视频数据进行测试,结果表明,本研究所提算法可有效地识别视频序列中的姿态,在每个动作仅包含 4 个训练数据的情况下,动作识别准确率均可达到 91.8%。

关键词:时序数据挖掘　动作识别　人体目标检测　人体姿态估计　动态时间规整

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2022)04-0700-08
DOI:10.13656/j.cnki.gxkx.20220919.010

随着视频监控网络的全面覆盖、移动互联网的不断普及、流媒体的逐渐兴起,产生了大量包含人体动作信息的视频数据。对视频数据中人体动作进行时序数据挖掘可用于监控安防、安全生产、人机交互、视频内容分析等方面,具有十分广泛的应用范围[1]。但是现阶段的动作识别算法需要大量的标记训练数据集,存在泛化性差、实时性差、场景受限的问题。

现有基于视频的动作识别算法主要分为 3 类:基于时空卷积的动作识别算法、基于双流卷积网络的动作识别算法以及基于人体骨骼[2,3]的动作识别算法。其中基于时空卷积的动作识别算法与基于双流卷积网络的动作识别算法直接利用时空卷积技术对视频帧流进行学习[4-11]。由于采用神经网络为学习框架,这类算法通常需要依赖大量的视频数据,且泛化性较差。基于人体骨骼的动作识别算法[12-14]利用人体姿态检测或专有设备提取人体的骨骼信息用于识别。

收稿日期:2022-03-21
【作者简介】
尹继尧(1981-),男,高级工程师,主要从事应急管理信息化与大数据分析研究,E-mail:Yinjiyao1@163.com。
【**通信作者】
【引用本文】
尹继尧,周琳,李强,等. 基于轻量化二维人体姿态估计的小样本动作识别算法[J]. 广西科学,2022,29(4):700-707.
YIN J Y,ZHOU L,LI Q,et al. A Small-sample Action Recognition Algorithm Based on Lightweight Two-dimensional Human Posture Estimation [J]. Guangxi Sciences,2022,29(4):700-707.

由于人体骨骼与背景无关,可以保证一定的泛化性,但是现阶段基于图卷积的骨骼动作分类同样需要一定的训练数据,且无法动态地扩展识别动作的类别。为此,本研究提出一种基于轻量化二维人体姿态估计的小样本动作识别算法,研究极少视频样本下多种动作的有效识别,并验证算法的有效性,以期降低动作识别算法对大规模数据的依赖。

1 相关工作

目前主流的3类动作识别方法中,基于时空卷积的动作识别算法如C3D[4]、I3D[5]、P3D[6]、T3D[7]、R2+1D[8]、SlowFast[9],以及基于双流卷积网络的动作识别算法如LSTM two-stream[10]、TSN[11]等,使用RGB图像、光流图像等像素级特征作为神经网络的输入,通过拟合训练实现动作的分类。但是这些方法会受到图像背景的干扰,泛化能力受限。基于人体骨骼的动作识别算法[15]相比其他算法更注重人体的信息,能够去除场景带来的干扰以适应更多的环境。现阶段主要采用基于图神经网络GCN架构的时空卷积模型进行训练[16,17],依旧需要一定量级的数据才能保证收敛。由于动作在空间与时间上存在歧义性与多样性,现有基于监督训练的方法普遍需要依赖大量的训练数据,这在实际应用中限制了算法的普适性[18]。因此本研究采用无需训练的方式来研究极少样本下多动作的有效识别,可以缓解动作识别任务对数据样本强依赖的现状,促进动作识别的落地。

此外,如何有效地从视频中获取和表征人体姿态信息是影响识别的关键。现有基于姿态估计的动作识别中的姿态信息主要来源于深度相机传感器标注和基于人体姿态估计提取。深度相机传感器虽然标注精准但是需要特殊的设备,硬件成本较高[12]。基于人体姿态估计的人体姿态表征虽然可以直接基于视频数据提取信息,但是由于需要多阶段的识别,需要权衡计算成本与识别精度[13,14]。因此,本研究同时研究轻量化二维人体姿态估计方法及与其配套的姿态动作特征构建方法,以保证在极少数据下动作识别的速度与准确性。

2 算法描述

本研究的算法如图1所示。该算法主要包括3个组件:轻量级人体检测算法HYOLOv5、基于Lite-HRNet[2]的二维人体姿态动作表征以及基于动态时间规整的小样本动作匹配。轻量级人体检测HYOLOv5基于小规模的YOLOv5算法,仅检测人体目标,能够有效地去除视频中与人体无关的背景信息。基于轻量化二维人体姿态估计Lite-HRNet的识别结果,算法根据动作的时空属性对人体姿态进行归一化表征,获取用于识别的姿态动作特征序列。考虑到仅使用极少样本进行识别,本研究采用模板匹配的思想,结合姿态动作特征序列特征设计姿态序列动态时间规整相似度度量方法,并通过类别中心选择算法降低匹配过程的时空复杂度,构建动作识别模板库用于动作识别。为验证算法的有效性,基于COCO 2017[3]构建Human COCO 2017数据集训练并测试HYOLOv5。本研究采集10种动作视频,在每个动作仅使用4个训练视频的情况下对算法进行测试。

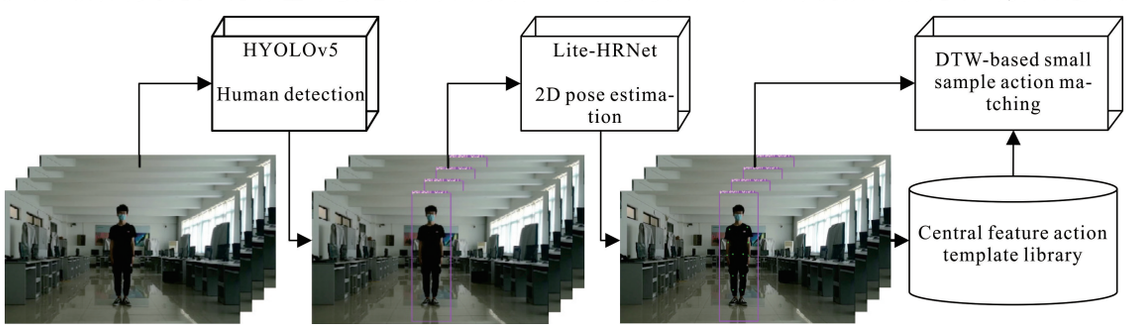


图1 算法示意图

Fig. 1 Schematic diagram of algorithm

2.1 轻量级人体检测器HYOLOv5

为了有效去除背景干扰,本研究构建轻量级的人体检测器。现有用于动作识别的人体检测算法通常是借助已经训练好的多目标检测器,通过类别过滤,仅保留人体检测框。然而这种方式会带来额外的计算成本,并且人体检测会受到其他类别信息的干扰,在与其他类别目标高度重合的时候会被误判为其他类别。因此本研究考虑使用已有的公开数据,重新训练仅用于识别人的目标检测器,进一步轻量化检测头。同时,考虑到动作识别的实时性要求,本研究最

终使用 YOLOv5-S 和 YOLOv5-N 作为骨干网络训练轻量级人体检测器 HYOLOv5。

YOLOv5 的核心思想是利用整张图作为网络的输入,直接回归边界框的位置坐标及其类别。具体的网络结构如图 2 所示,主要由 Backbone、Neck 和 Head 组成。Backbone 在输入端增加了 Focus 操作,即将输入图片等分切片成 4 份后堆叠,在不丢失信息的情况下将 RGB 通道扩充至 12 个,降低了网络运算的特征分辨率尺度。在 Darknet[19] 网络的基础上引入了 CSP[20] 结构来增强表征能力。Neck 层利用 CSP 结构构建特征金字塔(Feature Pyramid Networks,FPN),引入路径聚合网络[21](Path Aggregation Network,PAN)来对齐多尺度表征。

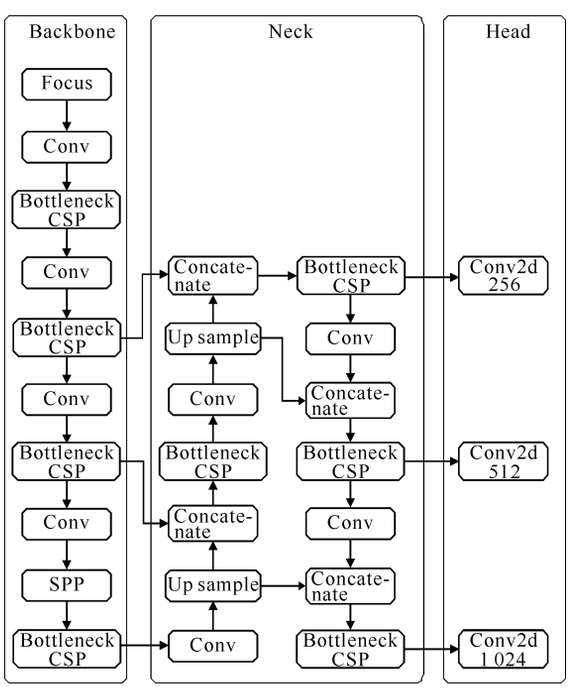


图 2 YOLOv5 模型结构

Fig. 2 Structure of YOLOv5 model

与 YOLOv5 用于多分类的 Head 不同,HYOLOv5 的类别为 1,因此网络的输出维度为 6,第 1 至第 4 维用于描述识别框,第 5 维为目标置信度,第 6 维为类别置信度。YOLOv5 设有深度系数与宽度系数来控制网络的规模,由小到大有 YOLOv5-N、YOLOv5-S、YOLOv5-M、YOLOv5-L 和 YOLOv5-X 5 种网络。HYOLOv5 同时在更大尺度上又提供了第 6 版系列权重,具有更高的准确率。

为训练 HYOLOv5,本研究提取了 COCO 2017 数据集中所有包含人标注的数据构建了 Human COCO 2017 数据集,使用原始训练集中的数据作为训练数据,使用验证集中的数据作为验证数据。依据迁移学习思想,基于 YOLOv5-S6 和 YOLOv5-N6 权重训练 HYOLOv5-S6 和 HYOLOv5-N6。与第 6 版系列权重输入分辨率 1 280 不同,为降低计算复杂度,HYOLOv5-S6 和 HYOLOv5-N6 的输入分辨率均为 640,模型的深度系数均为 0.33,宽度系数分别为 0.50 和 0.25。

经过极大值抑制算法即可对图像中的人进行目标检测。令检测到的人体框为$[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$,对应人体框的左、上、右、下边界。考虑到识别框会出现人体检测不全的情况,最终用于二维人体姿态估计的人体框描述数组(H)为

$$H=[x_{\min}-d_{\mathrm{l}}, y_{\min}-d_{\mathrm{t}}, x_{\max}+d_{\mathrm{r}}, y_{\max}+d_{\mathrm{b}}], \tag{1}$$

其中 d_{l}、d_{t}、d_{r}、d_{b} 分别为左、上、右、下边界的扩充像素数。

2.2 基于 Lite-HRNet 的二维人体姿态动作表征

在获取到人体框后,根据 H 从原始图像中裁剪出人体像素特征。对于之前的动作识别方法而言,人体像素特征可直接作为模型的输入特征进行训练,但是由于空间维度较大,往往需要一定的数据规模才能保证识别精度。因此,为了实现少样本数据下多动作的有效识别,本研究采用二维人体姿态信息作为人体动作表征的基础,其具有低空间维度与高行为描述的优势。

综合考虑识别精度与模型规模,本研究以轻量化二维姿态检测算法 Lite-HRNet 为基础,构建人体姿态特征描述算子。

Lite-HRNet 是 HRNet[22] 的轻量化版本。HRNet 的核心思想起源于 CPN[23] 工作中提到的:较高的空间分辨率有利于特征点精确定位,低分辨率具有更多的语义信息。为保证高分辨率特征的强度,采用网络并行连接从高到低的子网的方式来保持高分辨率表征,替代从低分辨率表征恢复高分辨率特征的方法,网络结构如图 3 所示。网络在设计中维持一个高分辨率表征的主干分支,在整个网络中不降低分辨率,为弥补高分辨率表征感受也受限的问题,并行引入渐进增加的低分辨率子网获取全局信息。同时,通过设计的特征融合模块来实现高、低分辨率表征的信息交换,用低分辨率信息增强高分辨率表征学习的同时,利用高分辨率表征获取的局部信息来增强全局的低分辨率表征。但是因为采用的是并行结构,且在骨干网络与特征融合模块大量使用高计算成本的卷积,参数的计算量很大。

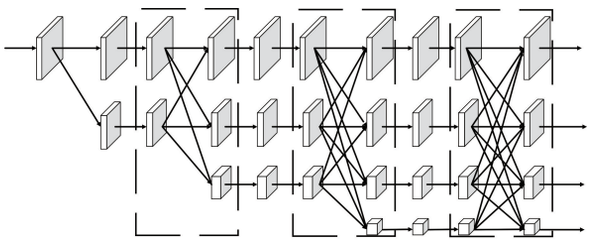


图 3　Lite-HRNet 网络结构

Fig. 3　Network structure of Lite-HRNet

为解决这个问题,Lite-HRNet 采用轻量化骨干网络 ShuffleNet[24] 的高效 Shuffle 块来替代 HRNet 中的基本模块。Shuffle 块的结构如图 4 所示。然而由于密集的平行子网间的信息交换,1×1 的卷积需要对每个 feature 的特征点进行遍历计算,成为计算的瓶颈。因此,通道加权(Conditional Channel Weighting,CCW)被提出来替代 1×1 的卷积,如图 4 所示。

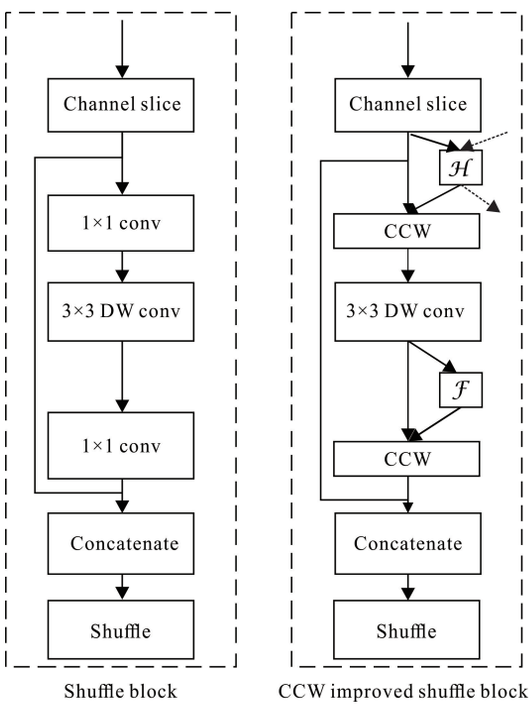


图 4　Lite-HRNet 基础模块结构

Fig. 4　Basic module structure of Lite-HRNet

Lite-HRNet 在 COCO 2017 验证数据集上根据网络深度与输入图像分辨率的不同提供了 4 种不同的预训练权重,如表 1 所示。由于二维人体姿态识别结果的精度与稳定性决定了动作识别的精度,本研究使用输入尺度为 384×288 的 Lite-HRNet-30 作为二维姿态特征提取网络。

表 1　Lite-HRNet 在 COCO 2017 上的结果

Table 1　Results of Lite-HRNet on COCO 2017

骨干网络 Backbone	输入尺度 Input scale	计算量 FLOPS	平均精度(%) AP (%)
Lite-HRNet-18	256×192	0.20	64.8
Lite-HRNet-18	384×288	0.45	67.6
Lite-HRNet-30	256×192	0.31	67.2
Lite-HRNet-30	384×288	0.70	70.4

在确定人体姿态特征后,需要进一步构建动作特征。令 Lite-HRNet 的识别结果为关节点坐标集合 P 与每个关节点对应的置信度 c,则

$$P=\{(J_{i,1},J_{i,2},\cdots,J_{i,17})\mid 1\leqslant i\leqslant t\},\quad(2)$$

其中 t 为总帧数,$J_{i,j}$ 为第 i 帧关节点 j 坐标(x,y),x 和 y 分别对应横、纵坐标,17 为 COCO 的关节点标注数。

对比每个关节点的置信度,发现“鼻子”“左眼”“右眼”“左耳”“右耳”(分别对应编号 1,2,3,4,5)的置信度不高,且存在大量闯动的情况,因此在构建人体姿态动作特征时不采用这 5 个点的信息。

每一个由二维人体姿态估计生成的关节点的坐标都是相对于 H 的绝对坐标,随着 H 坐标系的变化,关节点坐标的数值也会变化,因此需要坐标转换来获取与 H 无关的坐标描述。本研究选取每一帧的“左肩”和“右肩”的中心点 C 作为坐标原点进行坐标转换。由于人的体型、拍摄位置的影响,二维人体姿态估计生成的人体姿态在尺度上会有很大的差异,同样也会影响关节点的坐标,因此本研究使用初始帧中“左肩”与“右肩”的距离 D 作为人体姿态特征的标尺,经尺度归一化后获得人体姿态动作特征 A(如图 5 中红色虚线所示):

$$A=\{(\frac{J_{i,6}-C}{D},\frac{J_{i,7}-C}{D},\cdots,\frac{J_{i,17}-C}{D})\mid 1\leqslant i\leqslant t\}。\quad(3)$$

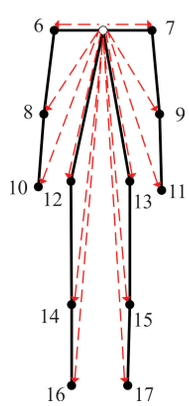


图 5　人体姿态动作特征

Fig. 5　Posture and movement characteristics of human body

2.3 基于动态时间规整的中心特征选择模板匹配

经过人体检测与姿态表征,高维视频序列被降维成低维姿态点集。基于深度学习的姿态行为识别,无论是监督、半监督或者自监督,通常需要一定量级的数据才能保证训练的精度,且识别的类别受限,无法满足极小样本下有效动作识别的需求。因此,本研究采用模板匹配的思想进行动作的识别。

为了有效度量两个人体姿态动作特征序列间的相似度,本研究提出了基于人体姿态动作特征的动态时间规整距离度量 ADTW。令人体姿态动作特征 A 的第 j 个关节点序列为 A_j,则

$$A_j=\{\frac{J_{i,j}-C}{D}\mid 1\leqslant i\leqslant t\}。\tag{4}$$

对于任意两个人体姿态动作特征序列 $A1$,$A2$,理论上可以直接计算 $A1_j$ 与 $A2_j$ 间的欧式距离来度量相似度。但是由于动作在时序上很难保证同步,且序列长度不一,因此本研究采用动态时间规整距离 DTW 来度量 $A1_j$ 与 $A2_j$ 间的相似性。通过对所有关节点序列的 DTW 值求和取平均,可以得到 ADTW 计算公式:

$$\mathrm{ADTW}(A1,A2)=\sum_{j=6}^{17}\mathrm{DTW}(A1_j,A2_j)/12。\tag{5}$$

基于 ADTW,根据少量多类动作视频来构建动作模板库。假设有 n 种动作,每种动作有 m 个训练数据,如果直接将对应的人体姿态动作特征存入动作特征库,直接利用 K-Nearest Neighbor (KNN)进行匹配分类,空间和时间复杂度至少为 $O(mn)$。并且如果录制过程中部分训练数据自身存在噪声,同样会影响动作识别的精度,因此本研究提出了基于类别中心选择的动作模板匹配方法,在新动作数据录入过程中动态选择每个动作中最具代表性的中心特征 Cent。

令 A^k 为某类动作第 k 个动作特征序列。计算 A^k 与所有类内动作特征序列的 ADTW 之和,用以度量该动作特征序列的重要性。所求的值越小,说明该动作特征序列与其他动作特征序列相比,与其他序列计算时获得更低 ADTW 值的可能性就越大,更能代表这个动作,则有

$$\min_i\sum_{j=0}^{m-1}\mathrm{ADTW}(A^i,A^j)\,,\tag{6}$$

中心特征 Cent 即为 A^i。动作特征库中仅存储每个类的中心特征,在匹配过程中复杂度降为 $O(n)$。

在构建完动作模板库后,动作的识别过程仅需计算待识别序列与每个类别的中心特征的 ADTW 距离,值最小的类别即为最终的识别结果。

3 验证实验

3.1 实验设置

实验采用的硬件实验环境为 Centos 7 系统,CPU 型号为 Intel Xeon Gold 5120 处理器,GPU 使用 2 张 NVIDIA GeForce 2080Ti,可用显存为 22 GB,使用 CUDA 10.0 与 Cudnn 7 进行深度学习加速训练,使用的深度学习框架为 Pytorch。

3.2 HYOLOv5 实验

如 2.1 节所述,本实验采用的数据集为 Human COCO 2017 数据集。数据集中共有 63 935 张训练集数据与 2 685 张测试数据。训练轮次为 300 轮,batch_size 为 64。考虑到轻量化需求,虽然采用了原分辨率为 1 280 的第 6 版系列权重,但是实际训练中的输入分辨率为 640。精度指标为识别精度,以及各类别在不同交并比下的平均准确率(mean Average Precision,mAP),主要有 mAP@0.5 和 mAP@0.5∶0.95。

为证明模型的优越性,算法在 Human COCO 2017 测试集上与 YOLOv5 原始权重进行对比,测试结果如表 2 所示,HYOLOv5 系列网络在识别的精度上均不弱于原始权重,且参数量低于原始权重,其中 HYOLOv5-S 的 mAP@0.5∶0.95 达到了 50.7%,在小规模人体检测网络中保持了较高的识别效果。

表 2 在 Human COCO 2017 上的识别结果

Table 2 Recognition results on Human COCO 2017

网络 Network	精度(%) Precision (%)	参数量(M) Parameter (M)	mAP@0.5 (%)	mAP@0.5∶0.95 (%)
YOLOv5-N6	80.0	3.2	68.3	40.1
YOLOv5-S6	82.3	12.6	74.6	46.9
HYOLOv5-N	80.0	1.8	75.0	45.9
HYOLOv5-S	83.3	7.0	78.9	50.7

3.3 动作识别实验

为验证小样本动作识别效果,本研究在不同室内环境下对多名体型各异的人员采集了 10 种肢体姿态的单人视频数据集,具体类别为侧抬右手、侧抬左手、侧推右手、侧推左手、右手上举、右手画 Λ、右高抬腿、左手上举、左手画 Λ 和左高抬腿,标签对应 0-9,每

个人员重复采集相同动作3－4次。结合实际应用情况，将每组动作的前4个动作序列作为训练集，剩下的作为测试集进行测试。训练与测试数据比例为1∶4，训练远少于测试数据。d_l、d_t、d_r、d_b的值均为60。

为证明基于动态时间规整的小样本动作匹配的有效性，利用相同数据使用KNN、Support Vector Machine(SVM)算法进行对比实验。实验结果如表3所示。经对比，在极少样本的情况下，KNN、SVM的识别精度远低于本研究的方法。在使用HYOLOv5-S作为人体检测器的情况下，本研究的方法在多类别分类上可以达到91.8%的准确率。从表中可以看出，人体检测器的精度会对动作识别的准确度造成影响。这说明对人体特征的有效表征能够降低视频动作识别对数据的强依赖，证明了小样本行为识别的可行性。

表3　动作识别结果

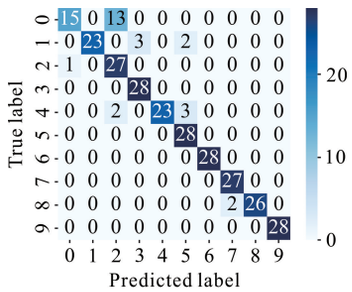
Table 3　Results of action recognition

方法 Method	检测器 Detector	准确率(%) Accuracy (%)
KNN	HYOLOv5-N	78.1
KNN	HYOLOv5-S	76.9
SVM	HYOLOv5-N	77.3
SVM	HYOLOv5-S	78.4
This study	HYOLOv5-N	90.7
This study	HYOLOv5-S	91.8

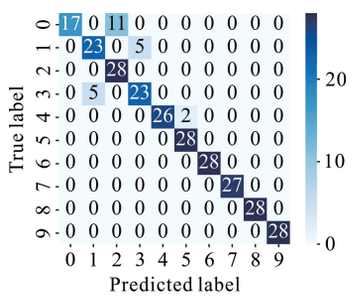
为进一步展示识别的细节，分别绘制了使用人体检测器HYOLOv5-N和HYOLOv5-S的动作分类识别混淆矩阵，如图6所示。识别的误判主要集中在存在细微差别的动作类上，如“侧抬右手”和“侧推右手”，但是在包含全身语义的动作中识别效果极佳，可达到100%的正确率。

4　结论

本研究提出了一种基于轻量化二维人体姿态估计的小样本动作识别算法，能够在极少视频样本下对多种动作进行有效识别。其中，轻量化二维人体姿态动作表征方法可以快速准确地提取视频中人体的特征，可以为其他基于姿态估计的动作识别算法提供数据基础。此外，用于动作识别的基于动态时间规整的中心特征选择模板匹配算法，为解决其他时序数据挖掘算法提供了思路。本研究的主要贡献包括4个方面：



(a) HYOLOv5-N



(b) HYOLOv5-S

图6　结果混淆矩阵

Fig. 6　Confusion matrix of results

①提出了一种基于轻量化二维人体姿态估计的小样本动作识别算法，仅需少量样本即可实现动作视频识别；

②构建了Human COCO 2017数据集并训练了轻量级人体检测算法HYOLOv5，可以有效地识别视频中的人体；

③基于轻量级人体姿态估计算法Lite-HRNet构建了人体姿态动作特征及姿态序列动态时间规整相似度度量方法；

④设计了一种基于中心特征选择的模板匹配算法，可以有效地降低模板匹配任务的时空复杂度，提高识别效率。

总体来看，本研究综合利用视觉智能算法将人的行为降维成时间序列表达，将行为识别问题简化建模为时间序列匹配问题，用灵活的识别机制来解决复杂的识别目标，具有一定的实际应用价值。未来的工作应包括2个方面：一是并行优化动态时间规整的运算效率，进一步提高算法的实时性；二是进一步提高轻量化人体姿态表征的精度以提升动作识别的精度。

参考文献

[1] 陈胜娣，魏维，何冰倩，等. 基于改进的深度卷积神经网络的人体动作识别方法[J]. 计算机应用研究，2019，36(3)：945-949，953.

[2] YU C Q，XIAO B，GAO C X，et al. Lite-HRNet：A lightweight high-resolution network [C]//Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021: 10440-10450.

[3] CHEN X, FANG H, LIN T, et al. Microsoft coco captions: Data collection and evaluation server [EB/OL]. [2022-02-23]. https://doi.org/10.48550/arXiv.1504.00325.

[4] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 4489-4497.

[5] CARREIRA J, ZISSERMAN A. Quo Vadis, action recognition? a new model and the kinetics dataset [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 6299-6308.

[6] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3d residual networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 5533-5541.

[7] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3d convnets: New architecture and transfer learning for video classification [EB/OL]. [2022-02-23]. https://doi.org/10.48550/arXiv.1711.08200.

[8] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 6450-6459.

[9] FEICHTENHOFER C, FAN H, MALIK J, et al. SlowFast networks for video recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 6202-6211.

[10] NG J Y-H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 4694-4702.

[11] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition [C]//European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, Cham, 2016: 20-36.

[12] DUAN H D, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE, 2022: 2969-2978.

[13] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 4724-4732.

[14] 刘勇,李杰,张建林,等. 基于深度学习的二维人体姿态估计研究进展[J]. 计算机工程,2021,47(3):1-16.

[15] 曾胜强,李琳. 基于姿态校正与姿态融合的 2D/3D 骨架动作识别方法[J]. 计算机应用研究,2022,39(3):900-905.

[16] YAN S J, XIONG Y J, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, LA, USA: AAAI Press, 2018: 7444-7452.

[17] 井望,李汪根,沈公仆,等. 轻量级多信息图卷积神经网络动作识别方法[J]. 计算机应用研究,2022,39(4):1247-1252.

[18] YANG X, LIU D, LIU J, et al. Follower: A novel self-deployable action recognition framework [J]. Sensors, 2021, 21(3): 950.

[19] REDMON J, FARHADI A. Yolov3: An incremental improvement [EB/OL]. [2022-02-23]. https://doi.org/10.48550/arXiv.1804.02767.

[20] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA: IEEE, 2020: 390-391.

[21] WANG C Y, XU C, WANG C H, et al. Perceptual adversarial networks for image-to-image transformation [J]. IEEE Transactions on Image Processing, 2018, 27(8): 4066-4079.

[22] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 5693-5703.

[23] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7103-7112.

[24] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 6848-6856.

A Small-sample Action Recognition Algorithm Based on Lightweight Two-dimensional Human Posture Estimation

YIN Jiyao[1], ZHOU Lin[1], LI Qiang[1], LIU Dongjingdian[2]
(1. Shenzhen Urban Public Safety and Technology Institute, Shenzhen, Guangdong, 518046, China; 2. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China)

Abstract: Action recognition is a research hotspot in temporal data mining in recent years with a wide range of application prospects. However, the action recognition algorithm based on deep learning at the present stage requires a large number of labeled training datasets, which has the problems of poor generalization, poor real-time performance and limited scene. In order to solve these problems, this study designs a small-sample action recognition algorithm based on lightweight two-dimensional human posture estimation. The algorithm builds a lightweight human detector HYOLOv5 based on the YOLOv5 algorithm. Based on the lightweight two-dimensional posture estimation model Lite-HRNet, a human posture feature descriptor is designed to effectively remove the interference of background on human action features. In order to effectively measure the similarity between temporal human posture feature descriptors, this study proposes a human posture feature distance measurement based on dynamic time warping, and designs an action template matching algorithm based on category center selection. The algorithm constructs a template library of action features through a small number of action videos, and uses the action template matching algorithm to achieve accurate recognition of multiple types of action videos. To verify the algorithm, this study tested HYOLOv5 on the COCO 2017 Humans dataset, and the human detection recognition accuracy of mAP@0.5∶0.95 could reach 50.7%. Based on 10 kinds of action video data, the results show that the proposed algorithm can effectively identify the posture in the video sequence. When each action contains only 4 training data, the accuracy of action recognition can reach 91.8%.

Key words: temporal data mining; action recognition; human target detection; human posture estimation; dynamic time warping

责任编辑:梁　晓