软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2015,26(Suppl.(2)):128−136

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

人体检测部分响应特征映射的人体动作识别*

王 磊, 吴 俊, 周志敏, 赵 旭, 刘允才

(上海交通大学 自动化系,上海 200240)

通讯作者: 赵旭, E-mail: zhaoxu@sjtu.edu.cn

摘 要: 在计算机视觉和多媒体领域,利用视觉信息进行语义层面人体运动分析非常重要且具有挑战性.提出一种利用检测信息的底层响应来描述人体动作的语义信息方法.在特定的人体动作下,可变形部分模型的检测结果隐含人体部分的关键信息,可以形成人体动作识别的特征.利用检测器的滤波器响应生成人体描述特征,对人体整体和部分的位置以及表观信息进行编码,由于该特征利用了人体部分相对于整体位置的统计信息,对检测过程中的误检部分具有较强的鲁棒性,基于该特征可将人体检测和动作识别融合成统一框架.在 3 个数据库上的实验结果显示了方法的有效性,取得了与其他方法相近或者更优的效果.

关键词: 人体检测;动作识别;可变形部分模型;特征提取;人体运动分析

中文引用格式: 王磊,吴俊,周志敏,赵旭,刘允才.人体检测部分响应特征映射的人体动作识别.软件学报,2015,26 (Suppl.(2)): 128−136. http://www.jos.org.cn/1000-9825/15023.htm

英文引用格式: Wang L, Wu J, Zhou ZM, Zhao X, Liu YC. Human action recognition through part-configured human detection response feature maps. Ruan Jian Xue Bao/Journal of Software, 2015,26(Suppl.(2)):128−136 (in Chinese). http://www.jos.org.cn/1000-9825/15023.htm

Human Action Recognition Through Part-Configured Human Detection Response Feature Maps

WANG Lei, WU Jun, ZHOU Zhi-Min, ZHAO Xu, LIU Yun-Cai

(Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: In computer vision and multimedia areas, it's an important yet challenging problem to perceive human motion at semantic level. In this work, a novel approach is presented to map the low-level response to semantic description of human actions. The features are based on the detection of deformable part models, in which the body pose information is contained implicitly under the specific human actions. The filter responses of the detectors are mapped to an effective feature description, which encodes the position and appearance information of human body and parts. The obtained features capture the relative configuration of body parts, and are robust to the false detections occurred in the individual part detectors. Comprehensive experiments conducted on three databases show the presented method achieves remarkable performance in most of the cases.

Key words: human detection; action recognition; deformable part model; feature formation; human motion analysis

语义层次的人体运动分析是可视媒体中一项重要研究内容,其研究目标是识别人体动作,并在此基础上实现对以人为中心的事件或环境的理解,可被广泛应用到智能监控、基于内容的检索、社交活动分析、人机交互、病人护理等领域.由于人体运动以及环境的复杂性,人体运动分析具有很大的挑战性,其主要问题是如何从底层可视媒体信息建立有效的描述特征,以及如何利用该特征在语义层次对人体运动进行建模和识别分析.

动作特征提取是动作识别的基础,研究人员采用的主要方法是对整个人体区域或者人体区域中随机选取的部分区域提取表观(appearance)特征[1],该类方法的缺点是提取的特征不能描述人体的姿态信息,但是姿态信

* 基金项目: 国家自然科学基金(61273285, 61375019); 国家重点基础研究发展计划(973)(2011CB30220)
收稿时间: 2014-06-20; 定稿时间: 2014-08-20

息在动作识别中能发挥重要的作用.最近的研究方法[2−4]利用人体姿态信息进行动作识别,取得了较好的效果,但同时也存在一些问题,例如该类方法都需要先对人体进行姿态估计,然后根据姿态估计的结果形成特征,该过程导致特征描述非常依赖人体姿态估计的结果,姿态估计的误差也会被引入到动作分析中.本文提出了一种新的方法,利用人体检测的信息从底层数据提取特征对人体动作进行建模,该方法可以在语义层次对人体动作进行描述,同时建模人体姿态信息和人体表观信息,最终把人体检测和动作识别融合成一个统一的框架.

可变形部分模型(deformable part model,简称 DPM)是由 Felzenszwalb 等人[5]提出的一种物体检测模型,该模型是一种由多个子模型构成的混合模型,其中每个子模型都是由一个根滤波器(root filter,描述人体整体信息)和若干个部分滤波器(part filter,描述人体部分信息)构成,即基于 DPM 的人体检测模型不仅包含人体的整体信息,还包括人体各个部分的表观以及位置信息.利用 DPM 进行人体检测可以获取根滤波器和部分滤波器的响应以及它们之间的相对位置关系,人体动作分析所需的姿态信息隐含在该响应和位置关系中,基于滤波器响应和位置关系能够形成有效的人体描述特征.基于模型的根滤波器和部分滤波器,人体的位置和表观信息是在检测过程中自动获取的,通过对人体检测区域的响应利用统计方法进行建模得到特征映射,而所获取的特征是基于响应的统计特征,因此对 DPM 中单个滤波器的误检有较强的鲁棒性.本文利用该方法建立对人体动作具有判别能力的描述模型,并使用该模型进行人体动作识别.

对比实验以及实验分析在3个数据库上进行:YouTube,HAT 和 Willow 动作数据库,包含多种复杂的动作类.本文方法的特征映射能够产生具有判别能力的动作描述子,在实验中取得了较好的结果,但同时还存在较大的改进空间.在实验结果和分析中,将详细展示和分析该方法的优缺点.

本文提出的方法是第 1 次尝试利用人体检测中的部分响应生成人体动作的语义描述,把人体检测和动作识别融合成统一的框架,构建了实现人体检测、人体姿态估计和动作识别的系统整体.

1 特征提取

本节介绍利用 DPM 进行人体检测的过程和基于该过程实现人体动作特征提取的实现细节.基于 DPM 的人体检测模型根据模型响应的大小实现人体检测,同时由于模型的根滤波器描述了人体的整体信息,而部分滤波器则描述了人体局部的位置和表观信息,因此对检测过程中两种滤波器的响应以及对应关系进行统计建模得到人体动作的描述特征.

1.1 检测过程

DPM 的实现细节见文献[5],本文的人体动作特征提取过程基于该模型的滤波器响应,因此本节对 DPM 方法的模型响应计算和检测过程进行简要描述.

假设一个 DPM 人体检测模型有 m 个子模型,每个子模型包含一个根滤波器和 n 个部分滤波器.利用改进的方向梯度直方图(histogram of oriented gradient,简称 HOG)作为特征,对一幅图像提取 HOG 特征金字塔 H,金字塔第 l 层坐标为 (x,y) 的位置用 $p=(x,y,l)$ 来表示.图像中人体所在的可能位置为 $z=(p_0,\ldots,p_n)$,z 描述了一个子模型的根滤波器和部分滤波器在 H 中的位置,p_0 表示根滤波器的位置,p_i 表示第 i 个部分滤波器的位置.模型的第 c 个子模型在位置 z 的响应值为每个滤波器的响应相加后,减去部分滤波器的变形损失,另外再加上一个响应的整体偏移:

$$\text{score}^c(p_0,\ldots,p_n)=\sum_{i=0}^{n}F_i'\cdot\phi(H,p_i)-\sum_{i=1}^{n}d_i\cdot\phi_d(dx_i,dy_i)+b \tag{1}$$

其中,$\phi(H,p_i)$ 是 H 中位置为 p_i 的图像特征向量,F_0' 是根滤波器形成的向量,F_i' 是第 i 个部分滤波器形成的向量,向量 $\phi_d(dx_i,dy_i)$ 表示部分滤波器的位置偏移变形,d_i 是变形损失系数,b 是响应的整体偏移.

根滤波器的位置决定图像中人体所在的位置.通过对各个部分滤波器进行变形处理得到根滤波器所在位置的子模型最大响应,该响应表示为

$$\text{score}^c(p_0)=\max_{p_1,\ldots,p_n}\text{score}^c(p_0,\ldots,p_n) \tag{2}$$

由于模型包含多个子模型, p_0 位置的最终响应是所有子模型中响应的最大值,即:

$$\text{score}(p_0)=\max_c \text{score}^c(p_0) \tag{3}$$

设置检测的阈值为 T,响应大于该阈值的位置构成了人体检测的目标集合 D:

$$D=\{p_0 \mid \text{score}(p_0)>T\} \tag{4}$$

为了去除被重复检测的目标,在 D 中应用非最大值抑制算法.把集合 D 中的数据按照响应从大到小排列,保证响应高的位置被优先选择,如果响应低的位置与响应高的位置重合面积超过 50%,那么该响应低的位置被忽略,即处于被抑制状态.重合面积的定义为

$$o=\frac{\text{area}(A)\cap\text{area}(B)}{\text{area}(A)} \tag{5}$$

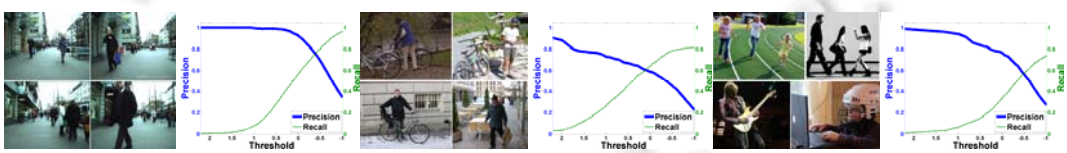
其中,area(A)表示在响应值较高位置的检测目标的区域面积,area(B)表示在响应较低位置的检测目标的区域面积, area(A)∩area(B)表示两者重合区域的面积.

为了评估检测结果是否正确,利用检测到的物体位置与物体真实位置(真值,手工标定)的重合度(overlap)来判断,重合度是根据两者覆盖区域的交集与并集的比值得到:

$$\text{overlap}=\frac{\text{area}(Det)\cap\text{area}(Gt)}{\text{area}(Det)\cup\text{area}(Gt)} \tag{6}$$

其中, area(Det) 表示检测位置的面积, area(Gt) 表示物体真实位置的面积,交集表示区域重合部分,并集表示两者覆盖的全部区域.本文采用 PASCAL[6]的标准,当重合度大于 0.5 时,认为检测结果正确.

为了测试模型在不同数据库上的检测效果,以及阈值对检测精度(precision)和召回率(recall)的影响,利用在 Pascal VOC 2010 [6]数据集上训练得到的模型分别在 3 个数据库测试:ETH 行人数据库[7]、INRIA 行人数据库[8]以及 Willow 动作数据库[9],检测结果如图 1 所示.子图中左列是 3 个数据库的示例图像,右列是对应数据库中检测的精度和召回率,其中横坐标是阈值,随着阈值的降低,精度逐渐下降,召回率逐渐提高.为了兼顾精度和召回率,一般选取阈值为–0.5 左右,对于较复杂的数据库,需要降低阈值以达到更高的召回率.



(a) ETH 行人数据库 (b) INRIA 行人数据库 (c) Willow 动作数据库

图 1 数据库示例图像以及对应的精度和召回率曲线

1.2 动作识别特征

人体各部分的位置表征了人体的姿态,人体部分的相对位置可以为动作识别提供判别信息,因此对人体部分的位置和人体的表观信息进行建模,形成动作识别特征.为了描述人体部分的位置信息,首先利用模型的根滤波器和部分滤波器检测到人体部分和整体的位置,通过两者的相对位置得到一个方向向量.对人体的表观信息,利用在检测过程提取的 HOG 特征,与根滤波器和部分滤波器进行卷积运算得到响应,把该响应作为人体的表观特征的描述.

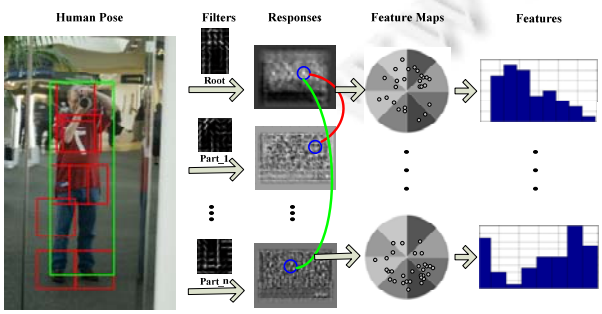


图2 基于检测的动作特征提取流程图

为了保证特征的鲁棒性,在检测到的人体各个部分和整体位置的邻域内对卷积响应进行采样,采样响应可以用来计算人体各个位置邻域内的统计特征.本文的动作特征是由人体的整体和部分的相对位置以及卷积响应共同决定的,特征提取的流程

图如图 2 所示,具体实现细节如下.

在人体检测过程中,得到检测结果的集合 D,通过公式(2)的计算过程可以得到由根滤波器和部分滤波器的位置构成的集合 Z:

$$Z=\{(p_0,...,p_n)\mid p_0\in D\} \tag{7}$$

其中, $p_i=(x_i,y_i,l_i)$ 表示 H 中第 l_i 层的 (x_i,y_i) 位置, p_0 表示根滤波器,其他为部分滤波器.

通过计算第 l_i 层的特征金字塔 H_i 与滤波器 F_i 的卷积得到响应:

$$R_i=H_i\otimes F_i,i=(0,...,n) \tag{8}$$

在响应 R_i 中以 s 为半径在(x_i,y_i)位置进行响应采样:

$$S_i=\{(x,y,r)\mid (x-x_i)^2+(y-y_i)^2\leqslant s^2\} \tag{9}$$

其中, r 是采样点的响应值.利用公式(9)得到根集合 S_0 和 n 个部分集合$(S_1,\ldots,S_n)$,出于分辨率的考虑,根滤波器和部分滤波器的计算在特征金字塔的不同层,因此需要把根集合中的坐标映射到部分集合所在的层,得到映射根集合 S_0',利用映射根集合和部分集合的相对关系和响应提取特征.

假设 (x^0,y^0,r^0) 是 S_0' 中的一个元素, (x^i,y^i,r^i) 是 S_i 中的一个元素,每个元素包含位置信息和模型响应,两者构成一个表征整体和部分的对应元素对 $\{(x^0,y^0,r^0),(x^i,y^i,r^i)\}$,其相对位置表示为

$$(x,y)=(x^i-x^0,y^i-y^0) \tag{10}$$

通过相对位置向量 (x,y) 计算得到向量的方向 $\theta(x,y)$,然后把方向离散化到 u 个区间,计算公式为

$$B=\text{round}\left(\frac{u\cdot\theta(x,y)}{2\pi}\right)\bmod u \tag{11}$$

每个元素对按照方向区间形成的直方图特征为

$$F_b^i=\begin{cases}r^0+r^i,\text{if } b=B\\0,\text{otherwise}\end{cases} \tag{12}$$

对 S_0' 和 S_i 中所有的元素对分别提取特征形成总体直方图统计特征 F^i,n 个部分集合得到最终的特征为 $F=(F^1,\ldots,F^n)$.

通过以上特征提取的过程看出,最终的动作特征表征了人体部分的相对位置关系以及人体整体的表观信息.在第 3 节中,利用支持向量机(support vector machine,简称 SVM)作为分类器,在 3 个动作数据库上进行人体检测和动作识别来测试动作特征的有效性.

2 实验分析

2.1 数据库

利用提出的动作特征在 3 个数据库上进行实验分析:(1) YouTube 视频数据库[10],(2) HAT(Human Attributes)数据库[11],(3) Willow 动作数据库[9].3 个数据库的示例图像如图 3 所示.原始的 YouTube 数据库没有动作类的标注信息,Ikizler-Cinbis 等人[12]对其中 11 个视频中的 775 幅图像进行了标注,包含 5 个动作类.HAT 数据库的主要目的是分析语义的人体属性,按照人的年龄、外观、姿态等分为 27 类.Willow 动作数据库是动作识别中较有挑战性的一个数据库,目的为了分析背景以及周围物体对动作识别的影响,一共有 7 类动作,分别是操作电脑、拍照、弹奏乐器、骑自行车、骑马、跑步以及行走.

2.2 实验设置

(1) 检测过程的参数选择.在实验中为了取得较高的精度和召回率,公式(4)中的阈值设置为–0.5.由于检测召回率的原因,通过检测所获得的数据比数据库中给出的标注数据少;由于精度的问题,数据中的部分负样本被误检为正样本,此类数据未参与动作分类.

图 3 数据库示例图像

(2) 分类器设置.利用 libSVM 工具包[13]来训练动作分类器.数据集被随机的分为训练集和测试集.支持向量机的核函数选用径向基函数(RBF),该核函数需要设置两个参数 C 和 γ.为了达到较高的准确率,采用交叉验证的方式选取参数.把训练集分为 5 组,用其中的 4 组训练,在另外一组上进行交叉验证,重复进行此过程,找到最优的参数组合.

(3) 性能度量.为了评估方法的性能,利用多种性能度量(performance metrics)来分析实验结果.比较常用的有准确率(accuracy)、精度(precision)、召回率(recall)和 F-score.

另外一个参数是总体准确率(overall accuracy),支持向量机的多类问题的准确率就是依据的该度量,总体准确率定义为 c/n,其中 c 是被正确分类的样本数目,n 是样本的总数.

文献[9]中对多类分类问题采用的度量为分类准确率(classification accuracy),通过计算混淆矩阵(confusion matrix)的对角线元素平均值得到,在本文实验中分类准确率也是每类召回率(per-class recall)的平均值.

2.3 实验结果

(1) 特征采样区间和采样半径.在第 2.2 节特征提取过程中,两个参数对特征的判别性能有较大影响:特征采样区间数目 b 和采样半径 s.b 决定特征的维数,s 反映特征所利用的响应范围.本文采用的分析方法为保持一个参数固定,然后分析另外一个参数对性能的影响.

b 决定直方图特征的维数,b 越大表示特征越细化,但同时特征也具有更高的维数.数据量一定的情况下,特征维数越高越容易产生过拟合的问题.如图 4 所示,其中左图为准确率,右图为 F-score.采样半径固定为 9,当采样区间为 9 时,动作识别的准确率和 F-score 最高.分别在 3 个数据库上对 b 进行测试,当 $b\in\{9,10\}$ 时,平均准确率和平均 F-score 较高.

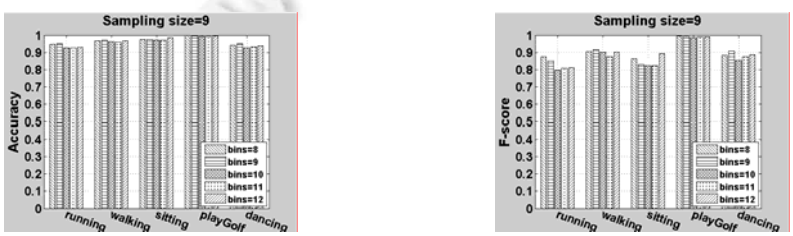


图4 YouTube数据库中不同采样区间数的准确率和F-score(采样半径为9)

特征是按照 s 在根滤波器和部分滤波器的检测位置进行响应采样,较小的采样半径能够更好地描述局部

特征,较大的采样半径包含较多的背景信息,不同的数据库在不同的采样半径下获取最好的效果.图 5 显示 YouTube 数据库的结果, 左图是准确率,右图是 *F*-score.其中采样区间数目固定为 9,当采样半径为 9 时,所有类的平均准确率和平均 *F*-score 最高.在其他的数据库上对采样半径的影响也进行了类似的实验,实验结果显示,当 $s\in\{7,8,9\}$ 时,各个动作类识别的平均准确率和 *F*-score 较高.

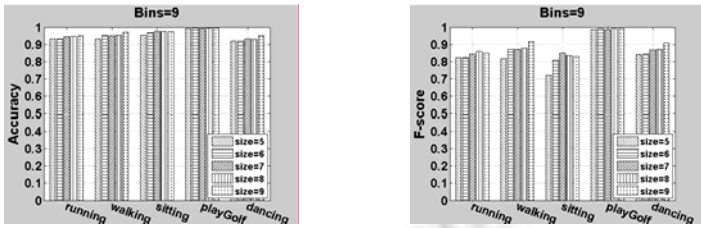


图5 YouTube数据库中不同采样半径下的准确率和*F*-score (区间数为9)

(2) HAT 数据库的动作识别结果.本文的方法侧重利用人体姿态信息进行动作识别,因此对 HAT 数据库中与姿态有关的动作类进行分析.表 1 列出了与姿态有关的 4 类动作的结果:站立(standing)、跑步或行走(runwalk)、坐(sitting)和下蹲(crouching).其中准确率较高的动作类是“坐”和“下蹲”,但该结果并不能充分说明这两类的分类效果好,因为 *F*-score 较低.产生该结果的主要原因是测试数据中正样本和负样本数目的不平衡,例如“下蹲”,负样本的数目与正样本的比例达到 40:1,准确率高较大程度受到负样本的影响,不能准确反映算法的性能,要综合考虑 *F*-score 的结果.在实验中,“站立”类别有 4 992 个样本,“跑步或行走”有 1 419 个样本,“坐”有 873 个样本,“下蹲”有 185 个样本.在 HAT 数据库中,部分数据含有多个类别的标注,导致分类更加复杂,本文方法在该数据集上达到的总体准确率为 64.95%.

表 1 HAT 数据库上的准确率和 *F*-score,其中采样半径为 7,采样区间数为 9

动作类别	所有数据集		平衡数据集	
	准确率	*F*-score	准确率	*F*-score
站立	0.643 1	0.771 7	0.761 2	0.548 4
跑步或行走	0.802 9	0.230 5	0.666 0	0.275 7
坐	0.867 6	0.096 8	0.687 4	0.441 6
下蹲	0.969 0	0.049 8	0.666 0	0.251 0

为了消除样本不平衡的影响,从原数据集中随机选取一个平衡的数据子集,所有动作类含有相同数目的样本.表 1 中平衡数据集的每一类都含有 129 个样本,从表格中可以看出“坐”和“下蹲”的 F-score 得到了明显的改善.性能最好的类别是“站立”,其中一个重要原因是“站立”的样本数目较多,训练效果较好.“跑步或行走”类别容易被误判为“站立”,“坐”容易被误判为“下蹲”,该结果也说明本文的特征能够有效的描述人体姿态,因为前两种动作表现在人体姿态上比较相近,而“坐”和“下蹲”的姿态也比较相似.与 Lazebnik 等人[14]和 Sharma 等人[15]等人的结果进行了比较,平均精度均值(mAP)结果见表 2.虽然本文的结果与其他的方法还有一定的差距,但是也具有一定的可比性,并且结果比其他方法差的一个重要的原因是其他方法的分类结果是基于标注真值的,而本文的结果是在检测识别统一框架下得到的.

表 2 HAT 数据库上平均精度均值(mAP)的对比,
其中本文的方法是检测和识别的综合

	SPM[14]	EPM[15]	本文的方法
mAP	0.555	0.587	0.321

(3) Willow 动作数据库的动作识别结果.Willow 动作数据库是为了分析背景、交互物体等对动作识别的影响,姿态信息在分类中的不是主导作用.本文方法在该数据库上没有取得很好的效果,原因是该方法侧重于利用姿态信息进行动作识别,分类准确率与 Delaitre 等人[9]的比较见表 3,本文方法的分类准确为 34.43%,而 Delaitre

等人的 LSVM 方法的分类准确率为 57.05%,LSVM+C2 方法的分类准确率更高,为 68.76%.在该数据库上,利用表观统计特征的分类效果要优于本文仅仅利用姿态信息的方法,本文方法的一个缺点是无法描述全局的表观统计特征,即不能对背景、交互物体进行建模.

表 3 在 Willow 动作数据上分类准确率的对比,其中本文的方法是检测和识别的综合

	lsvm[9]	lsvm+c2[9]	本文的方法
分类准确率	0.570 5	0.687 6	0.344 3

(4) YouTube 数据库的动作识别结果.Ikizler-Cinbis 等人[12]标注的 YouTube 数据库含有 5 类动作:跑步、行走、坐、打高尔夫球和跳舞,此类动作与人体姿态非常相关,与文献[12]的结果对比见表 4,本文方法的总体准确率提高了 16%.

表 4 YouTube 数据上总体准确率的比较

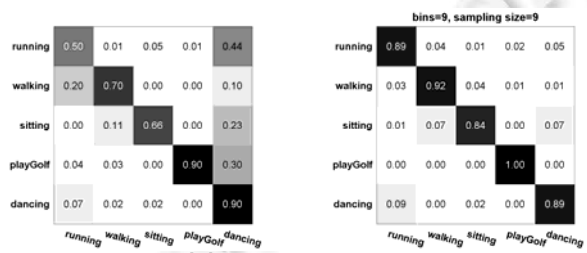
	Ikizler-Cinbis[12]	本文的方法
总体准确率	0.758 7	0.920 0

为了更细致的分析结果,在表 5 中列出了每一类动作的准确率、精度、召回率和 *F*-score,*F*-score 在“行走”、“打高尔夫球”和“跳舞”类别上均超过了 0.9,其他的类别也都大于 0.8.实验显示, “打高尔夫球”最容易被识别,该结果也与人们的直观感受保持一致,打高尔夫球的姿态较易于区分.

表 5 本文方法在 YouTube 数据库上各个动作类的准确率、精度、召回率和 *F*-score

动作类别	准确率	精度	召回率	*F*-score
跑步	0.948 9	0.823 6	0.886 6	0.851 6
行走	0.970 0	0.920 6	0.915 5	0.917 5
坐	0.974 4	0.826 9	0.844 0	0.829 9
打高尔夫球	0.995 6	0.989 8	0.996 0	0.992 8
跳舞	0.951 1	0.933 5	0.885 8	0.908 0

利用本文方法得到混淆矩阵与Ikizler-Cinbis等人结果的比较如图6所示,从图中能够看出他们的方法较明显的错误是把“跑步”混淆为“跳舞”,另外“跑步”与“行走”、“坐”与“跳舞”、“打高尔夫”与“跳舞”之间也比较容易发生混淆.利用本文的方法,结果有明显的改善,见图6的右图,数据都较好的集中在混淆矩阵的对角线上.



(a) Ikizler-Cinbis等人[12]的结果 (b) 本文的结果

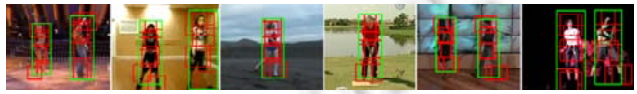
图 6 YouTube 数据库上混淆矩阵的对比

3 实验分析

本文利用人体检测过程中的信息生成人体姿态描述特征,并利用该特征进行动作识别,把人体检测和动作识别融合为一个统一框架.本文方法与其他方法有明显的不同.首先,大部分的方法是把人体检测和动作识别的过程分开设计,甚至不考虑检测结果,直接在人体标注真值区域中提取动作特征,但在实际应用中,必须考虑人体检测.其次,人体检测过程中的信息有助于动作识别,利用检测的中间信息形成动作特征能够有效减小复杂度,例如在本文的特征形成过程中只需要把检测响应统计成直方图特征,计算复杂度非常低.最后,基于本文检测识别框架的性能同时依赖于检测的准确率和识别特征的有效性,有较大的改进空间,例如改进人体检测模型

或者加入效果更好的动作特征,该框架为动作识别提供新的思路.

由第 2.3 节的结果可以看出,YouTube 数据库的结果要优于 HAT 和 Willow 动作数据库,利用 YouTube 和 Willow 动作数据库的不同进行详细分析.从图 1 的精度和召回率曲线可以看出,为了保证检测的准确率,Willow 动作数据库中有些复杂样本未被检测出,即检测的召回率低,而本文的检测识别框架是建立在检测中间结果基础上的,低召回率会影响整体的性能.YouTube 数据库的数据一致性比较高,检测的精度和召回率均比较高,检测过程不会对最终性能产生较大影响.图 7 显示了两个数据库的示例图像和检测结果,第 1 行是 YouTube 数据库的检测,第 2 行是 Willow 动作数据库的结果,由图可见 YouTube 数据库的检测比较精确,人体部分均被准确的包围在检测框内,而 Willow 动作数据库中的检测结果不是很理想,例如第 2 行图像(a)和(b),检测到的矩形框未能准确包围人体,并且检测框的位置和大小较不准确.检测过程中的此类问题会影响动作特征的形成,从而影响最终的动作分类精度.



(a) (b) (c) (d) (e) (f)

在YouTube数据库上的结果

(a) (b) (c) (d) (e) (f)

Willow动作数据库的结果

图 7 人体检测结果

两个数据库结果相差较大的另外一个原因是动作类别的不同,YouTube 数据库的动作类是“跑步”、“行走”、“坐”、“打高尔夫球”和“跳舞”,能够利用人体姿态信息进行区分,本文的方法恰好是侧重于描述人体的姿态.按照 Willow 动作数据库的设计,交互物体(如电脑、相机、自行车)在动作识别中发挥主要的作用,无法仅仅利用人体姿态信息来判别,这也是某些样本被误判为“行走”的原因,如图 7(c)中的“拍照”所示.

4 结 论

本文提出人体检测和动作识别的统一框架,在该系统中能同时实现人体检测、特征提取和动作识别,并取得了较好的效果.但同时由于本文的方法侧重于利用人体姿态信息进行动作特征描述和动作识别,还有较大的改进空间.在未来的工作里,将考虑在系统中加入交互物体(如马、自行车等)和全局表观特征的统计信息,以提高与交互物体有关的动作类别的识别性能.

References:

[1] Aggarwal JK, Ryoo MS. Human activity analysis: A review. ACM Computing Surveys, 2011,43(3):1−43. [doi: 10.1145/1922649.1922653]

[2] Zhao X, Liu Y, Fu Y. Exploring discriminative pose sub-patterns for effective action classification. In: Proc. of the ACM Multimedia. Barcelona: ACM, 2013. 273−282. [doi: 10.1145/2502081.2502094]

[3] Desai C, Ramanan D. Detecting actions, poses, and objects with relational phraselets. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, eds. Proc. of the European Conf. on Computer Vision. Florence: Springer-Verlag, 2012. 158−172. [doi: 10.1007/978-3-642-33765-9_12]

[4] Maji S, Bourdev L, Malik J. Action recognition from a distributed representation of pose and appearance. In: Proc. of the IEEE Computer Vision and Pattern Recognition. Providence: IEEE, 2011. 3177−3184. [doi: 10.1109/CVPR.2011.5995631]

[5] Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(9):1627−1645. [doi: 10.1109/ TPAMI.2009.167]

[6] Everingham M, van Gool L, Williams CKI, Winn J, Zisserman A. In: Proc. of the PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. 2010. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/

[7] Ess A, Leibe B, Schindler K, van Gool L. A mobile vision system for robust multi-person tracking. In: Proc. of the IEEE Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008. 1–8. [doi: 10.1109/CVPR.2008.4587581]

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the IEEE Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893. [doi: 10.1109/CVPR.2005.177]

[9] Delaitre V, Laptev I, Sivic J. Recognizing human actions in still images: A study of bag-of-features and part-based representations. In: Labrosse F, Zwiggelaar R, Liu YH, Tiddeman B, eds. Proc. of the British Machine Vision Conf. Aberystwyth: BMVA Press, 2010. 1–11. [doi: 10.5244/C.24.97]

[10] Niebles J, Han B, Ferencz A, Li FF. Extracting moving people from Internet videos. In: Forsyth D, Torr P, Zisserman A, eds. Proc. of the European Conf. on Computer Vision. Marseille: Springer-Verlag, 2008. 527–540. [doi: 10.1007/978-3-540-88693-8_39]

[11] Sharma G, Jurie F. Learning discriminative spatial representation for image classification. In: Hoey J, McKenna S, Trucco E, eds. Proc. of the British Machine Vision Conf. Dundee: BMVA Press, 2011. 1–11. [doi: 10.5244/C.25.6]

[12] Ikizler-Cinbis N, Cinbis RG, Sclaroff S. Learning actions from the Web. In: Proc. of the Int'l Conf. on Computer Vision. Kyoto: IEEE, 2009. 995–1002. [doi: 10.1109/ICCV.2009.5459368]

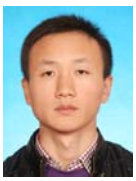
[13] Chang C, Lin C. LIBSVM: A library for support vector machines. ACM Trans. on Intelligent Systems and Technology, 2011, 2:1–27. [doi: 10.1145/1961189.1961199]

[14] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. New York: IEEE, 2006. 2169–2178. [doi: 10.1109/CVPR.2006.68]

[15] Sharma G, Jurie F, Schmid C. Expanded parts model for human attribute and action recognition in still images. In: Proc. of the IEEE Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 652–659. [doi: 10.1109/ CVPR.2013.90]

王磊(1986－),男,山东聊城人,博士生,主要研究领域为计算机视觉,模式识别.

吴俊(1991－),男,硕士生,主要研究领域为计算机视觉,模式识别.



周志敏(1989－),男,硕士生,主要研究领域为计算机视觉,模式识别.

赵旭(1974－),男,博士,副教授,主要研究领域为计算机视觉,模式识别.

刘允才(1948－),男,博士,教授,主要研究领域为计算机视觉,模式识别.