小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
2015 年 5 月 第 5 期
Vol. 36 No. 5 2015

一种结合姿态和场景的图像中人体行为分类方法

雷 庆[1,2,3,4],李绍滋[2],陈锻生[1]

[1](华侨大学 计算机学院,福建 厦门 361021)
[2](厦门大学 信息科学与技术学院智能科学与技术系,福建 厦门 361005)
[3](福建省大数据管理新技术与知识工程重点实验室,福建 泉州 362000)
[4](智能计算与信息处理福建省高等学校重点实验室,福建 泉州 362000)
E-mail: leiqing@hqu.edu.cn

摘 要: 提出一种结合姿态特征和场景信息对图像中的人体行为进行分类的方法,采用多尺度密集采样和 SIFT 特征对图像进行特征提取和描述,以非参数概率密度估计方法对特征空间的样本分布进行估计,并对概率密度梯度向量在码本单词上的聚集进行描述得到紧凑且有判别力的场景编码. 姿态分类则利用人体部位的表观和配置关系从图像中提取出与特定行为类别相关的姿态特征,利用最大分类间隔姿态分类器计算得到每个测试样本属于各个行为类别的评分值. 最后结合姿态分类器和行为场景分类器两种分类器输出值完成对测试样本的分类. 将本文的方法运用于 Willow-actions 数据集上进行评价,实验结果证明了该方法的有效性.

关 键 词: 人体行为识别; 姿态特征; 场景信息; 特征融合
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2015)05-1098-06

Using Poselet and Scene Information for Action Recognition in Still Images

LEI Qing[1,2,3,4], LI Shao-zi[2], CHEN Duan-sheng[1]

[1](College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)
[2](Cognitive Science Department, Xiamen University, Xiamen 361005, China)
[3](Fujian Provincial Key Laboratory of Data Intensive Computing, Quanzhou 362000, China)
[4](Key Laboratory of Intelligent Computing and Information Processing, Fujian Province University, Quanzhou 362000, China)

Abstract: This paper proposes a novel method for action recognition in still images using a combination of poselet and scene information. First, multi-scale dense sampling and SIFT descriptor are applied in feature extraction and description. Then non-parametric probability density estimation is employed to estimate the spatial distribution of feature space. To obtain discriminative scene feature, the gradient of probability density function is calculated and the vectors aggregation on visual words of action codebook is described for scene based action classification. While for pose classification, action-specific appearance and configuration patterns of human body part are extracted from training images, then a set of pose classifiers are trained to evaluate the class label confidence which test samples belongs to. Finally, the outputs of scene classifier and pose classifier are combined to decide the final class label. We validate our approach on Willow-action dataset and experimental results show that it achieves superior performance in comparison to several baseline methods.

Key words: human action recognition; pose feature; scene information; feature fusion

1 引 言

人体行为识别已经成为计算机视觉领域中一个重要的研究热点,近二十年的研究主要针对视频数据,其核心是采用计算机视觉和模式识别技术对图像序列进行分析、识别出人体的动作,并通过连续的跟踪以及上下文环境对行为进行推理和描述. 近年来由于在图像标注、基于内容的图像检索、视频压缩和人机接口等领域所具备的应用前景,静态图像中的人体行为识别研究也受到研究者们越来越多的关注. 受到视角和光照变化、遮挡、人体表观和姿态差异、以及复杂背景等因素的影响,图像中的人体行为识别方法仍然存在着难点和挑战.

目前的研究方法致力于从底层特征中建立有效的高层行为模型,运用鲁棒的识别算法完成对图像中人体行为的推理. 已有的方法主要分为整体模型、部位模型和交互模型三种,整体模型[1-4]从图像中检测出人体位置,但不需要对身体各个部位进行检测,利用整体的颜色、形状、轮廓等表观特征来表示动作; 部位模型[5-9]往往利用部位检测器得到的人体身体部位的位置,提取身体部位的表观以及部位配置关系所得到的姿态特征对行为进行表征; 而交互模型[10-13]主要针对交互行为中人-物体和人-人之间的共现位置和尺度关系对行为进

收稿日期: 2014-09-23 收修改稿日期: 2014-11-21 基金项目: 国家自然科学基金项目(61373076,61202143)资助; 高校博士学科点专项科研基金项目(20110121110024)资助; 中央高校基本科研业务费专项资金(11QZR04)资助; 福建省大数据重点实验室开放课题(2014KL03)资助; 华侨大学科研启动经费(13BS409)资助. 作者简介: 雷 庆,女,1980 年生,博士研究生,讲师,研究方向为人体行为识别、视频处理、计算机视觉; 李绍滋,男,1963 年生,博士,教授,博士生导师,研究方向为计算机视觉、人工智能; 陈锻生,男,1959 年生,博士,教授,研究方向为计算机视觉与模式识别、机器学习与数据挖掘.

行建模. 近年来主要采用姿态估计的方法通过提取人体部位表观和配置关系的姿态特征来识别图像中的人体行为,然而由于图像中的人体往往被遮挡导致无法有效检测出人体部位,此外身体部位配置关系所表达的语义与其所属的行为类别存在差距(如图 1 所示) ,使得仅仅依靠姿态估计的识别方法难以适应不同场景下的人体行为. 因此结合场景上下文提供的信息,通过对场景特征和姿态特征的提取,并建模出人体各部位以及与上下文环境的共现关系,有利于解决图像中人体行为尤其是交互行为的识别难题.

依据现有的研究基础,本文采用姿态识别和场景建模相结合的方法,研究了如何从图像中提取出有效的姿态特征对行为进行分类,如何通过场景信息的建模辅助得到行为的准确推理,提出了一种姿态分类和场景分类相融合的人体行为分类框架,如图 2 所示.

图 1　Willow-action 数据集 (Interacting with computer)

Fig. 1　Images from Willow-action dataset

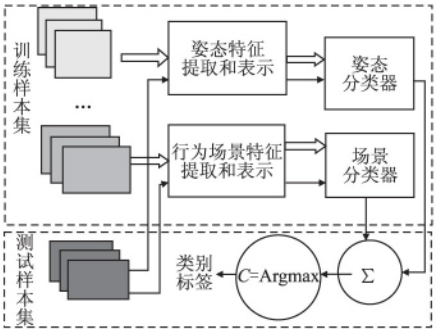

图 2　姿态和场景相结合的人体行为分类方法框架图

Fig. 2　Framework of our proposed approach

在该框架下,场景分类采用多尺度密集采样的特征提取和 SIFT 特征描述方式,利用基于高斯窗的核函数密度估计方法对特征空间样本点的概率分布进行估计,并计算特征点的概率密度梯度向量,通过统计梯度向量在 Kmeans 聚类后生成的码本上的聚集对特征空间中点集的分布形态进行描述. 姿态分类过程中利用部位检测器得到的图像中人体部位的位置,对部位的表观并提取部位之间的配置关系对行为进行表征. 采用最大分类间隔超平面学习方法为每个行为类别分别训练出场景分类器和姿态分类器,利用分类器输出的平均分值判定出测试样本所属的类别. 实验结果表明,本文的姿态特征提取以及与场景相融合的分类方法在图像行为识别基准数据集上取得了较高的识别准确率.

2　相关研究工作

胡琼等[14] 从行为特征提取、识别方法和评测数据库三个方面对已有的研究工作进行了总结,并阐述了目前行为识别领域存在的挑战. 谌先敢等[15] 提出了一种基于累积边缘图像的人体动作识别方法,首先利用形态学梯度操作获取人体轮廓,然后提取每一帧图像的形状特征并叠加起来得到累积边缘图并计算其 HOG 特征对人体行为进行分类. 吴娴等[16] 将视频中的人体行为看成由每帧轮廓图像组成的三位时空体,提出基于体积语义局部二值模式特征,并通过度量测试序列与训练样本之前的卡方距离对行为进行分类. 秦华标等[17] 提出一种将三维梯度方向直方图和光流直方图结合起来,并建立基于复合时空特征词典的人体行为识别算法对复合时空特征进行分类.

与本文的研究方法相似的是[5-7],Delaitre 等人[5] 首次提出结合身体部位特征和场景特征来表示行为,计算身体各个部位的表观特征,并采用 BOW 和空间金字塔相结合的特征编码方式对场景信息进行表征. Maji 等人[6] 提出了一种人体姿态和外观的分布表示,通过计算头部和躯干的三维朝向所得到的姿态激活向量来描述人体姿态. Zheng 等人[7] 采用[6]中提出的姿态激活向量对姿态进行分类; 关于场景分类则采用结合 SIFT 特征表示和稀疏编码方法,将图像映射到超完备词典所表示的坐标系下的特征空间. 而本文提出的姿态与场景相结合的分类方法,从姿态和场景特征的提取、特征编码方式以及分类方法都与[7]不同,比较如表 1 所示.

表 1　与其他研究工作的比较

Table 1　Comparison of our method with other research works

方法	特征提取		特征编码	分类方法
	姿态特征	场景特征		
Delaitre[5]	形变部位模型	尺度不变特征	词袋模型	非线性支持向量机
Maji[6]	姿态激活向量	—	—	线性支持向量机
Zheng[7]	姿态激活向量	梯度方向直方图	稀疏编码	最小化二次误差
本文	部位表观与配置关系	尺度不变特征	梯度向量聚集	线性支持向量机

3　场景特征表示和分类方法

场景特征的提取和表示以多尺度密集间隔采样的方法从图像中提取出特征点,底层特征运用 SIFT 局部特征描述方法对特征点计算得到 128 维的梯度方向直方图描述子. 针对 BOW 的缺点,本文根据[18]中基于核函数的非参数概率密度估计思想,利用概率密度的梯度向量在行为码本上的聚集所描述的特征点分布形态对图像进行编码,算法框图如图 3 所示.

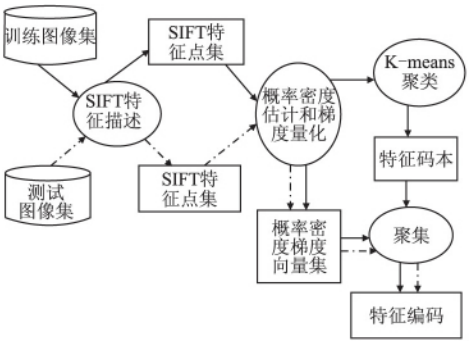

图 3　场景特征提取和表示框图

Fig. 3　Framework of scene feature extraction and description

3.1　基于概率密度梯度向量的场景特征提取

传统的 BOW 方法简单地统计子空间的特征出现频度所

得到的聚类中心直方图作为图像特征的编码,往往导致较大的量化误差且无法得到特征空间点集分布形态的有效表示.以描述特征点在特征空间中的分布形态为目的,本文采用非参数概率密度估计方法对特征点的分布进行估计,利用特征点的概率密度梯度向量及其在码本单词上的聚集形态进行描述,并运用主成分分析方法抽取出一组具有判别力的特征,将原始特征映射到以新的特征作为坐标基的新特征空间中.

假设从一幅图像 I 中提取出的 SIFT 描述子集合为 $x=\{x_1,x_2,\cdots,x_n\}$,利用公式(1)的高斯窗非参数估计方法对图像中特征点的分布进行估计:

$$\hat{p}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right) \tag{1}$$

其中 n 表示从图像中采样出的 SIFT 特征点数,x_i 表示第 i 个 128 维的 SIFT 特征描述子,σ 表示非参数估计中所采用的高斯窗参数(实验中对该参数的选取进行了评估).

求公式 1 对 x 的导数,进一步计算出特征点的概率密度梯度向量,如公式(2)所示:

$$\nabla\hat{p}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma^3}(x_i-x)\exp\left(-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right) \tag{2}$$

运用 PCA 主成分分析方法,从公式 2 计算得到的训练样本特征点梯度向量集合 $V=\{v_1,v_2,\cdots,v_n\}$ 中提取出一组有判别力的特征向量 $U=\{\mu_1,\mu_2,\cdots,\mu_d\}$,将这一组正交的基向量作为新的坐标系,将原始空间的特征点梯度向量投影到新的坐标系下,即通过公式 U^TV 计算出梯度向量在这组新的正交基上的投影坐标,得到新的梯度向量集 $V'=\{v_1',v_2',\cdots,v_n'\}$,其中 $v_i'=[\mu_1^Tv_i,\mu_2^Tv_i,\cdots,\mu_d^Tv_i]$.

3.2 基于梯度向量聚集和显著性表示的场景特征编码

在特征空间运用 Kmeans 算法对特征点集进行聚类生成特征码本(实验中选择码本的维数为 256,该维数在实验章节中进行了评估和说明).根据显著性表示的思想[19],通过计算特征与最近单词之间的距离,与到距离其他 s-1 个最近的单词之间距离的差异所反映出的单词显著度来建模特征对该单词的贡献度.

假设从训练图像样本集合中提取出的 SIFT 描述子集合为 $X=\{x_1,x_2,\cdots,x_N\}$,运用 Kmeans 算法聚类得到 k 个单词 $W=\{w_1,w_2,\cdots,w_k\}$,采用欧式距离 $d(x,y)=\|x-y\|_2^2$ 量化特征点到单词的距离值,通过公式(3)计算出特征对该单词的贡献度所示:

$$p(w_i\mid x)=\sum_{j=2}^{s}\left(\|x-\tilde{w}_j\|_2-\|x-\tilde{w}_1\|_2\right)/\|x-\tilde{w}_j\|_2 \tag{3}$$

其中 $\tilde{w}_j$ 表示按从小到大的顺序对距离 x 最近的 s 个单词进行排序后,距离 x 最近的第 j 个单词.

对每一幅图像 I,以每个单词为中心统计梯度特征向量集 $V'=\{v_1',v_2',\cdots,v_n'\}$ 在 k 个单词 $W=\{w_1,w_2,\cdots,w_k\}$ 上的聚集编码:

$$s_i=\sum_{j=1}^{n}p(w_i\mid x_j)\|v_j\|_2v_j' \tag{4}$$

最后将 k 个单词的梯度聚集向量串联起来,得到图像的特征编码表示 $s_I=\{s_1,s_2,\cdots,s_k\}$.

3.3 场景分类

场景分类采用多对一的支持向量机分类方法(1-vs-all SVM),采用线性核函数的样本距离度量方法:

$$K(s_i,s_j)=s_i^Ts_j \tag{5}$$

为行为类别集合 $C=\{c_1,c_2,\cdots,c_q\}$ 中的每一种行为类别训练出一个最优分类超平面:

$$\min_{w,b}\frac{1}{2}\|w\|^2 \tag{6}$$

$$s.t.\ y_i\ [w^Ts_i+b]-1\geqslant 0\qquad i=1,\cdots,N \tag{7}$$

其中,y_i 表示图像编码对应的行为类别标签,s_i 表示训练样本中第 i 个图像编码,w 和 b 分别表示分类超平面的法向量和阈值权.

对每一种行为类别,将属于该行为类别的图像编码 s_i 作为正样本,其他类别图像编码作为负样本,训练得到一个最大间隔分类超平面.对于测试样本 $\tilde{x}$ 的分类决策方法则是将分类器输出值看作样本的类别隶属度,选取最大的得分值作为测试样本的类别:

$$\hat{c}=\arg\max_{c\in C}(w_c^T\tilde{x}+b_c) \tag{8}$$

4 姿态特征表示和分类方法

基于部位模型的人体检测方法[20]通过选择合适的特征描述子构建出人体部位的检测器,同时利用部位之间的几何关系构建出合理的部位配置关系模型对人体位置进行推理.这类方法将人体分为几个部位并建立每个部位的表观模型、以及部位与部位之间的位置关系模型,并提出评价图像中待检测位置与模板之间匹配程度的能量函数,通过最大化匹配程度(或最小化表观和形变程度)的求解过程推理出最可能的人体位置.

近年来,基于部位模型的姿态估计方法同样流行于图像中的人体行为识别.在部位检测输出结果基础上,针对每一种行为类别的训练样本,提取出每个部位的表观特征及其配置

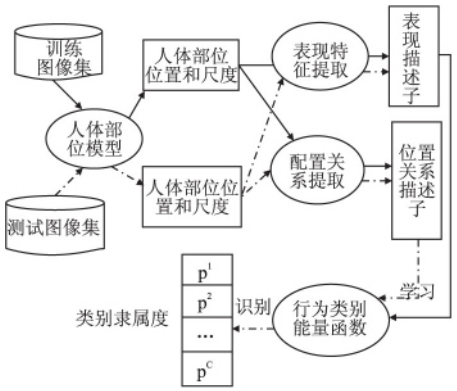

图 4 姿态特征提取和行为识别框图

Fig. 4 Framework of pose classification

信息,为每一种行为类别定义一个能量函数,利用最大间隔分类超平面的方法从正负训练样本中学习出该行为类别的能量函数参数,根据测试样本在不同行为类别的能量函数上计算得到的评分值,选取最大分值所代表的类别隶属度完成对测试样本的分类,具体算法框图如图 4 所示.

4.1 人体部位检测

由于行为识别需要更多的部位表观和位置变化等细节信息,因此为了提取出更丰富的姿态特征,本文采用[21]提出的人体关节模型,该模型将人体分为 26 个部位(头部 2,左躯干 4,右躯干 4,左手 4,右手 4,左腿 4,右腿 4),利用模型在图像上进行部位检测的输出结果(如图 5 所示),提取出与特定行为类别相关的部位表观和配置关系特征.

图 5　人体部位模型[21]在 Willow-actions 上的检测结果

Fig. 5　Detection results on Willow-actions by parts model[21]

4.2　姿态特征提取和表示

一个包含 n 个部位的形变部位模型表示为 $P=(P_0,P_2,\cdots,P_{n-1})$,其中 P_i 表示第 i 个部位的部位检测器. 每个部位检测器可用四元组 (F_i,d_i,v_i,σ_i) 来表示,其中 F_i 表示第 i 个部位表观特征的权向量,d_i 是一个四维的度量部件位置的权向量(二次函数的系数),表示每个部位相对于其理想位置的形变代价. v_i 是一个二维向量,表示第 i 个部位相对于整体检测器的位置. σ_i 是表示部件尺度的数值,每个部位 P_i 的位置表示为 $p_i=(x_i,y_i,\sigma_i)$.

姿态特征的提取包含部位表观 $F(p_i)$ 和部位形变特征 $\phi_d(p_i)$ 两部分,其中部位的表观特征采用 HOG 特征描述方式对每个部位检测结果计算出表观特征描述子; 而形变特征的提取首先以头部位置 $p_0=(x_0,y_0,\sigma_0)$ 作为固定点,根据公式(9)计算出其他部位的尺度归一化空间位移量:

$$(dx_i,dy_i)=\left(\frac{x_i-x_0}{\sigma_0},\frac{y_i-y_0}{\sigma_0}\right)-v_i \tag{9}$$

部位的形变代价采用部位相对位移所造成损失的二次函数进行度量,因此形变特征用公式(10)的 4 维向量表示:

$$\phi_d(dx_i,dy_i)=(dx_i,dy_i,dx_i^2,dy_i^2) \tag{10}$$

4.3　参数学习和姿态分类过程

对于给定行为类别的姿态模型参数 $w=(m_0,m_1,\cdots,m_{n-1},d_1,d_2,\cdots,d_{k-1},b)$,图像中的部位检测结果 $l=(p_0,p_1,\cdots,p_{n-1})$ 与姿态模型的匹配程度度量通过最大化公式(11)所示的能量函数来实现,该能量函数衡量了各部位表观的相似性以及形变的代价:

$$score(l)=\sum_{i=0}^{n-1}m_iF(p_i)-\sum_{i=1}^{n-1}d_i\phi_d(dx_i,dy_i)+b \tag{11}$$

其中第一项可以看作姿态特征在表观滤波器上的响应值,提取部位的 HOG 特征后计算与姿态模型的响应值,越大则与该类行为越匹配. 第二项表示部位的相对位移在特定行为类别模型下所造成的损失,损失越小则表示与该类行为越相似,b 是利用最大间隔分类超平面的参数学习方法中引入的偏置项.

令特定行为类别的姿态模型参数为 $w=(m_0,m_1,\cdots,m_{n-1},d_1,d_2,\cdots,d_{n-1},b)$,姿态特征表示为 $\Phi(l)=(F(p_0),F(p_1),\cdots,F(p_{n-1}),-\phi_d(p_1),-\phi_d(p_2),\cdots,-\phi_d(p_{n-1}),1)$,那么公式 11 的能量函数形式化为: $score(l)=w\cdot\Phi(l)$.

利用 1-vs-all 的多类线性 SVM 方法为每个行为类别的姿态模型学习出一个最优的分类超平面:

$$L(w)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\max(0,1-y_iw\Phi(x_i)) \tag{12}$$

对于给定的测试样本 $\tilde{x}$ 的姿态分类决策方法与场景决策相同,将分类器输出值看作样本的类别隶属度,选取最大的得分值作为测试样本的类别.

5　实　验

5.1　Willow-actions 数据集

Delaitre 等人[5]从 Flickr 网站上采集了用户上传的 911 张图像建立了 Willow-actions 数据集,包含了丰富和自然的视角、姿态和着装等变化,同时存在着不同情况的遮挡且往往以复杂的场景作为背景. 按照 Pascal VOC 的标准以人工的方式对每张图像中的人体位置和行为类别进行了标注. 该图像数据集包括七个行为类别 "Interacting with computers", "Photographing", "Playing a musical instrument", "Riding bike", "Riding horse", "Running"和"Walking",每个类别的图像数量不少于 109 张,其中除"Riding bike"行为类别的图像来自于 Pascal 2007 VOC 数据集以外,其他类别的图像均来自于 Flickr 的关键字搜索后经过人工筛选后的图像集. 图像集合被分为训练集合和测试集合,每个行为类别的 70 张图像作为分类模型的参数进行训练,剩下的图像作为测试样本对分类方法的性能进行评价.

5.2　参数设置和场景分类实验结果

场景分类过程中需要设置的参数包括: 特征点采样尺度和步长、Kmeans 聚类单词数、概率密度估计中高斯核方差、概率密度梯度投影正交基数量,以及显著性编码中选取的特征点近邻数. 经过反复实验选取出效果最佳的参数配置,其中 SIFT 特征点的采样尺度 patchsize 设置为[16,24,32],得到 16×16,24×24 和 32×32 三种尺度大小的图像块,采样间隔 step 设为 8 个像素点; 高斯窗概率密度估计中选取核函数的方差 σ^2 设置为 512^2(262144); 训练过程中 SIFT 特征点的 Kmeans 聚类中心点的个数 k 设置为 256; 同时梯度向量投影坐标系中正交基数量为 256; 显著性编码过程中选取特征点周围的 10 个中心对特征点属于单词的权重进行计算.

按照[5]的三种实验方案并增加一种按照图像中的人体标注 bounding box 裁剪出仅包含人体的图像区域共四种方案进行实验,分别表示为: a) 图像、b) 人体、c) 人体(1.5 resale)、d) 人体 + 图像,其中实验方案 a 表示在不知道人体位置区域的前提下直接从整幅图像中提取特征进行分类; 实验方案 b 表示利用数据集中人工标注好的每一幅图像中的人体位置区域,从背景去除后的人体区域图像中提取出特征进行分类; 实验方案 c 表示按照图像中人体位置,从图像中裁剪出 1.5 倍大小的图像区域进行分类中. 实验方案 d 表示对人体区域和整幅图像单独进行特征提取,在线性 SVM 分类器的学习中将两种特征的核矩阵叠加起来,即 $K(x,y)=K_p(x,y)+K_b(x,y)$; 本文的方法在 Willow-actions 数据集上得到的分类准确率如下页表 2 所示.

由表 2 可以看出由于数据集中存在的视角、姿态和着装变化、以及复杂背景信息的影响,在引入人体位置信息的先验知识下(实验方案 b,c,d),七种行为类别的分类效果好于对

整幅图像进行分类(实验方案 a).仅仅依靠人体图像而不使用任何背景信息的情况下,只在 running 行为类别上取得了最好的效果,而结合人体图像并引入适当的背景信息在 interacting with computers 和 riding bike 行为类别上取得了最好的分类效果,而将人体特征和背景信息分别进行特征提取以及相似性度量,将两种特征对应的相似度矩阵叠加起来进行分类在 photographing,playing music,riding horse 和 walking 行为类别上取得了最好的效果.本文的方法在 Willow-actions 数据集上进行四种方案的实验,最后得到的平均分类准确率分别为:58.49%,59.05%,60.33% 和 63.4%.四种方案对应的混淆矩阵如图 6 所示.

表 2 四种实验方案在 Willow-actions 上的场景分类效果

Table 2 Scene classification performance for the four methods on Willow-actions

行为类别/实验方案	a) 图像	b) 人体	c) 人体(1.5 倍)	d) 人体+图像
Interacting with Computers	82.05%	56.41%	**94.87%**	61.54%
Photographing	24.68%	19.48%	7.79%	**27.27%**
Playing Music	63.56%	71.19%	52.54%	**77.97%**
Riding Bike	78.42%	83.45%	**84.17%**	82.73%
Riding Horse	73.68%	71.93%	73.68%	**84.21%**
Running	38.27%	**60.49%**	55.56%	55.56%
Walking	48.76%	50.41%	53.72%	**54.55%**
平均分类准确率	58.49%	59.05%	60.33%	**63.40%**

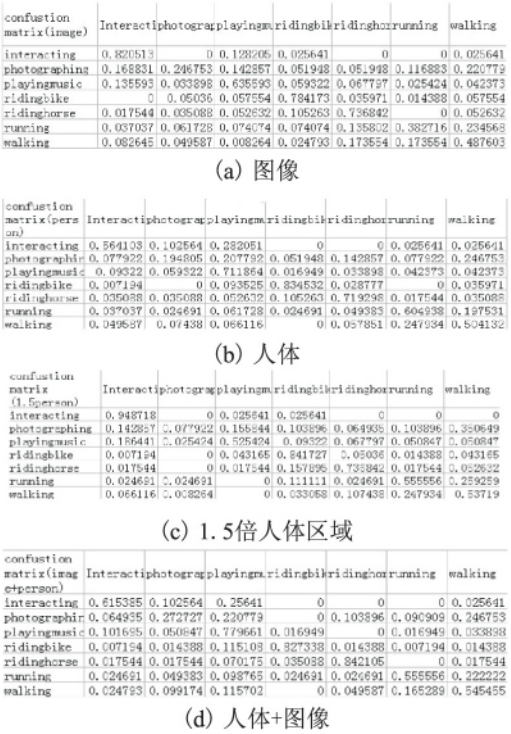

图 6 四种实验方案下场景分类的混淆矩阵

Fig. 6 Confusion matrixes of scence classification for the four different methods

在 a,c,d 三种实验方案下所得到的分类准确率与 [5] 进行对比,如表 3 所示,可以看到在三种方案下本文的场景分类方法所得到的准确率比 [5] 有较大的提升.然而与 [5] 所不同的是,本文的方法在方案 d 中即裁剪出包含少量场景信息的人体图像区域进行分类,比方案 c 将人体和图像的核矩阵叠加后进行分类的效果更好.

表 3 场景分类与 [5] 的实验结果对比

Table 3 Comparison of our scece classification performance with [5]

实验方案/方法	Delaitre et. al [5]	本文的方法
a) 图像	53.97%	58.49%
b) 人体	—	59.05%
c) 人体(1.5 rescale)	55.9%	60.33%
d) 人体+图像	58.86%	63.4%

5.3 姿态与场景分类融合的实验结果

场景和姿态的特征融合有三种方案:第一种是在特征表示层面,将提取出的场景特征和姿态特征串联起来作为最终的图像特征表示,在统一的 SVM 框架下去求解最优的行为分类超平面;第二种是在分类器的训练过程中,将场景特征和姿态特征对应的核函数矩阵叠加起来归一化后求解,这种方式下如果选择线性核函数作为特征之间相似度的度量则得到的效果与第一种方案相同;第三种方案是对场景特征和姿态特征分别训练分类器,将分类器的输出分值看作样本对应于各个行为类别的隶属度,将两种分类器的输出分值结合起来取得分最高的类别作为样本分类结果.经过对比实验发现,第三种特征融合的方案下所得到的分类效果最佳,表 4 中列出了其在 Willow-actions 数据集上的最终分类结果.

表 4 Willow-actions 上两种分类器相结合的分类准确率

Table 4 Overall classification performance combined pose with scence on Willow-action

行为类别/实验方案	图像+姿态	人体(1.5 倍)姿态	人体+图像+姿态
Interacting with Computers	79.05%	69.23%	82.05%
Photographing	39.87%	25.56%	31.69%
Playing Music	65.25%	67.97%	59.32%
Riding Bike	74.82%	85.61%	85.27%
Riding Horse	77.19%	82.46%	77.86%
Running	44.44%	45.68%	64.20%
Walking	48.80%	58.68%	58.55%
平均分类准确率	61.06%	62.17%	65.56%

从表 4 可以看到,本文的方法在 5.2 的 a,c,d 三种实验方案下得到的平均分类准确率分别为:61.06%,62.17%,65.56%,姿态与场景特征相结合在 photographing,riding biking,running,walking 行为上的分类效果得到了较大的提升,原因可能在于 photographing,running,walking 这几种行为所对应的场景存在很大的视角、光照和背景变化,场景特征容易产生混淆,结合姿态特征进行分类将有助于提升分类效果.

最后将本文的方法与已有的类似研究方法 [5,7] 进行了比较,从下页表 5 中可以看出本文的方法在 Willow-actions 数据集上取得了更好的分类效果.

5.4　码本数量对场景分类效果的影响

首先在码本中心个数的选择上,选取了 K 设置为 32,64,128,256 和 512 进行了实验,图 7 给出了四种实验方案下码本中心个数的选择对场景分类效果的影响. 从图 7 中可以看到随着聚类中心数量的增加,平均分类准确率也有所提高,当中心数增加到 256 时 a,b,c 三种方案的分类准确率分别达到最高的 58.49%,59.05% 和 60.33%,实验方案 d 的最好分类效果 63.4% 在选取 K 为 128 时获得; 然而随着码本中心数增加到 512,四种实验方案分类准确率都出现下降,其中方案 d 下降得最快,原因可能在于实验方案 d 结合人体区域和整幅图像的特征对行为进行分类,码本中心数量太大造成不同行为下人体和背景特征点分布形态的过拟合,使得码本的判别能力大大降低.

表 5　本文方法与其他方法的比较

Table 5　Comparison of our method with other works

实验方案/方法	Delaitre et. al [5]	Zheng et. al [7]	本文的方法
图像 + 姿态	—	—	61.06%
人体 + 姿态	—	—	62.17%
人体 + 图像 + 姿态	62.15%	65.40%	65.56%

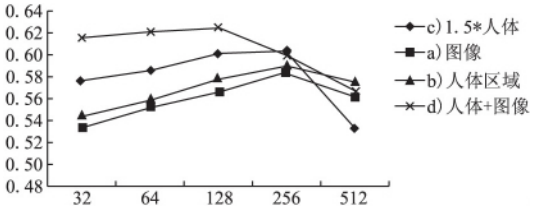

图 7　码本数量对场景分类效果的影响

Fig. 7　Scene classification performance on various numbers of codewords

6　结论及未来工作

本文提出了一种结合姿态和场景信息对图像中的人体行为进行分类的方法,采用局部概率密度梯度向量聚集和显著性编码相结合的表示方法对行为发生的场景特征进行描述,并训练出 SVM 分类器对行为发生的场景进行分类; 此外利用部位检测器得到的图像中人体部位位置,提取出与特定行为类别相关的部位表观和配置关系特征,同样采用最优分类超平面的方法从样本中训练得到特定类别的姿态分类器; 结合两种分类器的输出分值所表示的类别隶属度对样本的类别进行决策. 本文的方法在 Willow-actions 数据集上进行了实验,结果证明了方法的有效性. 未来的工作致力于将本文的方法应用于视频数据,通过对视频中的场景建模、并从连续图像帧中提取出与姿态变化相关的特征表示人体行为.

References:

[1] Li P, Ma J, Gao S. Actions in still web images: visualization, detection and retrieval [C]. Web-Age Information Management: 12th International Conference, vol6897, Springer, 2011: 302-313.

[2] Li P, Ma J. What is happening in a still picture? [C]. IEEE First Asian Conference on Pattern Recognition, 2011: 32-36.

[3] Wang Y, Jiang H, Drew M, et al. Unsupervised discovery of action classes [C], IEEE Conference on Computer Vision and Pattern Recognition, 2006: 1654-1661.

[4] Thurau C, Hlavac V. Pose primitive based human action recognition in videos or still images [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[5] Delaitre V, Laptev I, Sivic J. Recognizing human actions in still images: a study of bag-of-features and part-based representations [C]. Proceedings of the British Machine Vision Conference, BMVA, 2010, 97.1-97.11.

[6] Maji S, Bourdev L, Malik J. Action recognition from a distributed representation of pose and appearance [C] IEEE Conference on Computer Vision and Pattern Recognition, 2011: 3177-3184.

[7] Zheng Y, Zhang Y J, Li X. Action recognition in still images using a combination of human pose and context information [C]. IEEE International Conference on Image Processing, 2012: 785-788.

[8] Sharma G, Jurie F, Schmid C. Expanded parts model for human attribute and action recognition in still images [C]. IEEE Conference on Computer Vision Pattern Recognition, 2013: 652-659.

[9] Bourdev L, Malik J. Poselets: body part detectors trained using 3d human pose annotations [C]. IEEE International Conference on Computer Vision, 2009: 1365-1372.

[10] Shapovalova N, Gong W, Pedersoli M, et al. On importance of interactions and context in human action recognition [C]. Pattern Recognition and Image Analysis: 5th Iberian Conference, Springer, 2011, 6669: 58-65.

[11] Yao B, Fei-Fei L. Modeling mutual context of object and human pose in human-object interaction activities [C]. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010: 17-24.

[12] Prest A, Schmid C, Ferrari V. Weakly supervised learning of interactions between humans and objects [C]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2012, 34: 601-614.

[13] Prest A, Schmid C, Ferrari V. Weakly supervised learning of interactions between humans and objects [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3): 601-614.

[14] Hu Qiong, Qin Lei, Huang Qing-ming. A survey on visual human action recognition [J]. Chinese Journal Computers, 2013, 36(12): 2512-2524.

[15] Shen Xian-gan, Liu Juan, Gao Zhi-yong, et al. Recognizing realistic human actions using accumulative edge image [J]. Acta Automatica Sinica, 2012, 38(8): 1380-1384.

[16] Wu Xian, Lai Jian-huang. Human action recognition based on volume semantic local binary patterns [J]. Chinese Journal Computers, 2012, 35(12): 2652-2660.

[17] Qin Hua-biao, Zhang Ya-ning, Cai Jing-jing. Human action recognition based on composite spatio-temporal features [J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26(8): 1320-1325.

[18] Kobayashi T. BoF meets HOG: feature extraction based on histograms of oriented p. d. f gradients for image classification [C] Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013: 747-754.

[19] Huang Yong-zhen, Huang Kai-qi, Yu Yi-nan, et al. Salient coding for image classification [C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011: 1753 -1760.

[20] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester: cascade object detection with deformable part models [C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010: 2241-2248.

[21] Yi Yang, Deva Ramanan. Articulated pose estimation with flexible mixtures of parts [C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011: 1385-1392.

附中文参考文献:

[14] 胡　琼,秦　磊,黄庆明. 基于视觉的人体动作识别综述 [J]. 计算机学报,2013,36(12): 2512-2524.

[15] 谌先敢,刘　娟,高智勇,等. 基于累积边缘图像的现实人体动作识别 [J]. 自动化学报,2012,38(8): 1380-1384.

[16] 吴　娴,赖剑煌. 基于体积语义局部二值模式的行为识别 [J]. 计算机学报,2012,35(12): 2652-2660.

[17] 秦华标,张亚宁,蔡静静. 基于复合时空特征的人体行为识别方法 [J]. 计算机辅助设计与图形学学报,2014,26(8): 1320-1325.