高技术通讯 2013 年 第23 卷 第12 期:1258 ~1267

doi:10.3772/j.issn.1002-0470.2013.12.008

基于特征联合和直方图交叉核函数的动作识别方法①

张世辉②* ** 高文静* 孔令富* **

(*燕山大学信息科学与工程学院 秦皇岛 066004)

(**河北省计算机虚拟技术与系统集成重点实验室 秦皇岛 066004)

摘 要 为提高动作识别的识别率和实时性,提出了一种新颖的基于特征联合和直方图交叉核函数的动作识别方法。该方法首先跟踪视频中运动物体上的局部时空特征点形成运动轨迹,并计算出轨迹的梯度方向直方图(HOG)、光流直方图(HOF)、运动边界直方图(MBH)特征和轨迹上各点所在视频帧局部区域的局部二值模式(LBP)特征组成联合特征矩阵;然后等量地对每种动作的各训练样本的联合特征矩阵进行平均采样,将采样结果合并后运用 bag-of-features 方法进行 K-means 聚类形成码书,在此基础上利用码书量化各样本的联合特征矩阵得到表示视频样本中运动信息及结构信息的特征向量;最后将形成的特征向量作为支持向量机(SVM)的输入,同时选择直方图交叉核函数作为 SVM 的核函数,训练动作识别的分类器并进行测试。实验结果表明,该方法不仅提高了动作识别的识别率,而且通过利用直方图交叉核函数可缩短分类器的训练与测试时间。

关键词 动作识别,运动轨迹,联合特征,bag-of-features,直方图交叉核函数

0 引 言

近年来动作识别已经成为计算机视觉领域中的研究热点,研究成果普遍应用于人机交互、活动监督、体育事件分析、视频索引及修复等方面。目前已有的动作识别方法大体分为三类。一是基于时序或状态模型的动作识别,该类方法通过对视频中运动人体的检测、分割与跟踪,提取人体轮廓特征并建立动作的时序或状态模型,然后比较输入视频中相应特征与模型的相似性以达到动作识别的目的[1-4],其计算简单、方便快捷,但是识别结果易受视角变换和部分遮挡的影响。二是基于上下文及语境分析的动作识别,该类方法将人体运动信息与全局中的场景或物体编码为一个整体进行动作识别[5,6],其识别效果较好,但计算量大且只适用于特定的场景。三是基于局部时空特征点的动作识别,该类方法将视频序列看成是时间域与空间域上的三维时空卷,然后在三维时空卷上探测局部时空特征点并通过计算其特征进行动作识别[7-9],其计算复杂性低且对噪声和部分遮挡不敏感,具有一定的鲁棒性,但要求特征点足够稠密,以便能够保证涵盖运动目标的全部信息。

基于局部时空特征点的动作识别方法简单高效,近年来引起了学者们的广泛关注。如何在保证特征点稠密的基础上描述时空特征点的特征是影响人体动作识别方法效果的关键因素。现有水平下比较著名的时空特征点特征有 Scovanner[10] 提出的 3D-SIFT 特征、Klaser[11] 提出的 HOG3D 特征和 Bay[12] 提出的 SURF 特征等。然而,由于单一特征往往受到人体外观、环境、摄像机设置等因素的影响从而使其适用性受到了限制。为此,Wang[13] 通过跟踪运动的时空特征点组成稠密轨迹并提取轨迹的梯度方向直方图(histogram of oriented gradient, HOG)、光流直方图(histogram of optical flow, HOF)和运动边界直方图(motion boundary histogram, MBH)特征组成联合特征,然后对训练样本的联合特征随机采样,结合 bag-of-features[14] 方法进行动作识别。该方法利用光流法检测视频序列中所有运动着的点,去除了严格意义上的检测、分割,减小了计算的复杂性,同时由于 HOG 特征对运动目标的光学

① 863 计划(2006AA04Z212),国家自然科学基金(61379065)和河北省自然科学基金(F2010001276)资助项目。

② 男,1973 年生,博士,教授;研究方向:视觉信息处理,智能并联机器人,模式识别等;联系人,E-mail:sshhzz@ysu.edu.cn

(收稿日期:2013-05-14)

变换和几何变换有较强的鲁棒性以及 HOF 和 MBH 特征对局部运动和相对运动信息能够进行较为准确的描述,从而明显提高了识别率,遗憾的是此方法没有考虑局部特征点所在视频样本帧的纹理信息,使得某些运动相近的动作容易发生混淆,降低了识别率;在结合 bag-of-features 方法进行特征处理时,是对全部训练样本的轨迹联合特征进行笼统的随机采样,导致各动作间参与聚类的特征行数的比例不平衡,使得实验结果较难重现;在分类器核函数的选取上存在不足。为了解决上述问题,本文提出了一种新颖的基于特征联合和直方图交叉核函数的动作识别方法,该方法将运动轨迹的形状、运动信息与视频帧局部的纹理特征相结合形成联合特征,经过 bag-of-features 方法进行平均采样后,输入选用直方图交叉核函数的支持向量机(SVM)进行分类器的训练与测试。实验表明,与已有方法相比,所提方法既有效地提高了动作识别的识别率,同时也降低了分类器的训练与测试时间。

1 方法描述

本文提出的动作识别方法分为训练和测试两个阶段,如图 1 所示。训练阶段利用训练样本集训练出动作识别的分类器。首先,跟踪训练样本中运动物体上的特征点形成运动轨迹,并计算出轨迹的 HOG、HOF、MBH 及局部二值模式(local binary pattern,LBP)特征形成联合特征矩阵。其次,对训练样本的特征矩阵按动作种类数目平均采样后进行 K-means 聚类形成码书。然后,利用码书量化所有样本的特征矩阵,将各样本的特征矩阵转化成由每个单词所得票数形成的特征向量。最后,将各样本对应的特征向量作为 SVM 的输入训练动作识别分类器得到分类器模型。测试阶段对测试样本进行测试。对每一个测试样本按照训练阶段的方法依次进行特征提取和特征量化处理,然后利用训练阶段所得分类器模型进行分类测试,最终得到动作识别结果。为了更加清晰地论述本文所提动作识别方法,下面将动作识别分为特征提取、特征处理以及分类器的训练与测试三个过程进行详细论述。

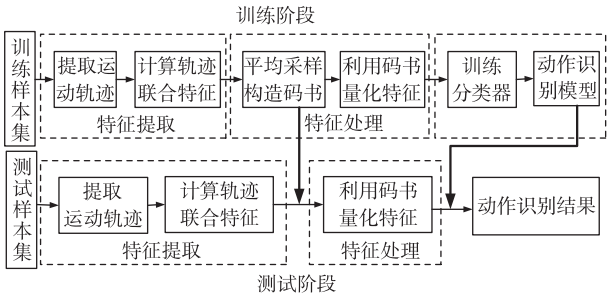


图 1 方法流程图

1.1 特征提取

1.1.1 提取时空特征点形成稠密轨迹

目前已有较多的时空特征点检测方法。Gunnar[15]利用改进的光流法进行时空特征点检测,具有计算简单且准确率高的特点,故本文采用该方法完成特征点的检测。基于检测出的特征点即可形成稠密轨迹,本文借鉴文献[13]的方法形成稠密轨迹,具体过程是先对视频中的每帧图像建立 8 个尺度空间图像序列,比例因子为 $1/\sqrt{2}$,然后为确保特征点的稠密性,在每帧的每个尺度空间上对检测到的特征点分别每隔 5 个像素点进行采样。在此基础上对采样后的各时空特征点施加全局平滑约束函数,根据中值滤波原理跟踪到下一帧相应尺度空间上该特征点的位置,持续进行跟踪操作直到累计 15 帧时,判断是否成功跟踪到 15 个特征点。若是,则形成由 15 个特征点串联成的轨迹,否则,去除该跟踪过程中已跟踪到的特征点。对视频每帧的各尺度空间图像序列循环上述检测与跟踪过程直至视频结束即可形成稠密轨迹。最后,计算轨迹上各特征点位置的标准差。若标准差小于给定阈值(即出现轨迹静止情况)或大于给定阈值(即出现轨迹跳变情况),则抛弃该轨迹,否则存储该轨迹相应信息以便进行下一步操作。图 2 是从某一视频的第 t 帧、第 $t+1$ 帧和第 $t+2$ 帧上提取到的时空特征点及分别连续跟踪各时空特征点 M 帧后到达第 $t+M$ 帧、第 $t+M+1$ 帧和第 $t+M+2$ 帧时的运动轨迹示意图。

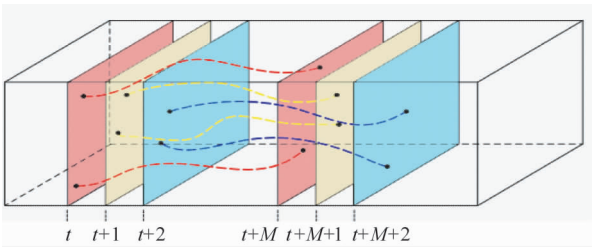


图 2 视频帧中时空特征点和运动轨迹的示意图

1.1.2 计算轨迹特征

提取出特征点形成运动轨迹后,下一步需要计算出轨迹的相应特征。现有用于动作识别的特征,HOG、HOF 和 MBH 皆表现出很好的性能。HOG 特征聚焦于局部目标的表观和形状信息,HOF 特征捕

获了局部的运动信息,MBH 特征则着重于局部的相对运动信息,此三种特征的结合,全面描述出了轨迹的形状和运动趋势。但是,为了突出运动轨迹与所在视频帧的相对关系,本文在 HOG、HOF、MBH 三种特征的基础上又结合了运动轨迹所在每一帧视频图像的 LBP 特征,形成了一个完整的动作识别特征描述体系。下面介绍这几种特征的计算方式。

为了计算运动轨迹的 HOG、HOF 和 MBH 三种直方图特征,首先需要为运动轨迹建立一个以轨迹为中心,大小为 $N \times N$ 像素、M 帧的时空卷。类似于图像的直方图特征计算方法,需要为这个时空卷进行分块,块的大小为 $n \times n$ 像素、m 帧。然后根据运动角度的范围确定特征通道数后,依次计算每个小块中各像素的相应特征,将特征值按照划分好的通道进行投票。最后将各小块的投票结果按固定顺序首尾串联作为最终的直方图特征结果。特征维度的计算公式为

$$f = \frac{N \times N}{n \times n} \times \frac{M}{m} \times num_{bin} \tag{1}$$

其中,N 为时空卷的长和宽,M 为时空卷的帧长,n 为块的长和宽,m 为块的帧长,num_{bin} 为各直方图特征的通道数。根据经验,本文取 $N = 32$,$M = 15$,$n = 16$,$m = 5$。对 HOG 特征与 MBH 特征,设置 $num_{bin} = 8$,由于 HOF 特征要把 0 作为单独通道,所以 HOF 特征的 $num_{bin} = 9$。综上,$f_{HOG} = 96$,$f_{HOF} = 108$。本文中 MBH 特征的计算方法是将光流场分为水平和垂直方向上的两个分量,然后分别计算两个分量上的光流梯度,所以,$f_{MBH-X} = 96$,$f_{MBH-Y} = 96$ 分别代表 MBH 在 X 方向上的特征维度与 Y 方向上的特征维度。

不同于以上三种直方图特征,LBP 特征是一种用于表示图像纹理信息的特征算子,是典型的结构与统计相结合的纹理分析方法。该算子通过将图像中某一像素点与其相邻像素点的灰度值进行比较并阈值化比较结果来描述该像素点所在位置的纹理变化模式。由于该算子在计算相邻像素点的信息时既简单又高效且具有光照不变性,有利于快速捕捉到视频中有运动发生时各帧的纹理信息,所以本文利用 LBP 特征来分析运动轨迹所在视频帧间的局部纹理变化。LBP 特征提取过程如下所述。

首先将时空卷中 $N \times N$ 像素平面映射到视频帧上,定位时空特征点所在视频帧上的局部区域,将此局部区域的图像转换成灰度图像。然后遍历灰度图像中的各像素点的灰度值,并将其与其八邻域内像素点的灰度值进行比较。最后通过公式

$$LBP(c) = \sum_{n=0}^{7} f(g_n, g_c) \cdot 2^n,$$

$$f(g_n, g_c) = \begin{cases} 1, & g_n \geqslant g_c \\ 0, & g_n < g_c \end{cases} \tag{2}$$

计算出运动轨迹上所有点所在局部区域的 LBP 特征。式中,c 为图像中除边界点外的任一像素点,g_c 为 c 点的灰度值,g_n ($n = 0,1,2,\cdots,7$) 为点 c 自左上像素点起沿顺时针方向八邻域像素点的灰度值。按此方法计算每个像素点的 LBP 特征值后,建立 8 通道统计直方图并对结果进行归一化,最终形成维度为 $f_{LBP} = 8 \times 15 = 120$ 的特征向量。此过程记录了以运动轨迹上所有点为中心的视频帧局部区域的纹理信息。提取出 LBP 特征后,与已得到的 HOG、HOF 和 MBH 特征首尾衔接进行特征联合即可组成 516 维的联合特征矩阵,该矩阵中每一行的 516 维特征就描述了一条运动轨迹的形状、运动及其所在视频帧的局部纹理信息。

1.2 特征处理

1.2.1 平均采样构造码书

通常情况下,在一个视频序列中提取到的运动轨迹数以千计乃至上万,这使得特征矩阵十分庞大而且在特征联合过程中容易产生噪声,因此,为了能简单高效地利用已提取到的特征,需要对特征进行处理。基本思想是将前述提取到的联合特征矩阵按照 bag-of-features 方法进行 K-means 聚类形成码书。需要特别指出的是为了减少计算的复杂性,一些文献在进行聚类时并没有采用全部的特征,而是随机选取特征矩阵中的某些行进行聚类[16]。这使得实验结果时好时坏、随机性较大,实验结果的可再现性明显降低。同时,随机选取特征也使得复杂动作(或者运动轨迹数量多的动作)在码书构造过程中的贡献大,简单动作(或者运动轨迹数量少的动作)在码书构造过程中的贡献小,从而容易导致将简单动作误识为复杂动作。所以,本文在进行大量实验后,总结出一种既能保证识别率又能尽量实现结果唯一性的码书构造方式:首先确定参加聚类的总的轨迹条数 N_T(一般取 10^5 数量级),然后用(N_T/动作种类数)计算出各种动作参与聚类的轨迹条数的平均值 N_A,最后依据 N_A 分别在每种动作的各训练样本的特征矩阵中平均抽取(N_A/每种动作的训练样本数)行轨迹特征并将其按行合并后作为 K-means 聚类的输入,聚类结果即为码书。

由于本文中联合特征矩阵由 HOG、HOF、MBH

和 LBP 四种特征组成,同大多数文献一样,我们将每种特征聚类单词的个数设置为 4000 个,所以码书的大小为 16000。为便于理解,给出了图 3 所示的平均采样构造码书过程的示意图。

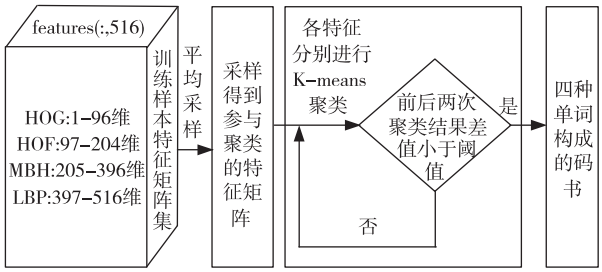


图 3　平均采样构造码书过程示意图

1.2.2　量化特征矩阵

码书构建完成后,通过将代表各视频样本的特征矩阵对码书进行投票处理即可得到各视频样本量化后的特征向量。投票方式是分别计算视频样本特征矩阵中每条轨迹的四种特征与码书中相应种类单词间的欧氏距离,每条轨迹中的每种特征为与它距离最近的相应单词投一票,即每条轨迹投出四票。投票结束后,根据 16000 个单词分别得到的票数建立统计直方图,所得结果即为每个视频样本对应的特征向量,该向量的维度为 1 × 16000。

图 4 展示了部分视频样本对应的特征向量示意图。其中,第 1 – 4000 维代表与 HOG 特征对应的 4000 个单词分别获得的票数,依此类推,第 4001 – 8000 维、第 8001 – 12000 维、第 12001 – 16000 维分别代表与 HOF、MBH、LBP 特征对应的 4000 个单词分别获得的票数。

BOW <600x16000 double　BOW <600x16000 double　BOW <600x16000 double　BOW <600x16000 double>

	1	2		4001	4002		8001	8002		12001	12002	12003
1	0	0	1	0	11	1	0	0	1	9	0	163
2	10	1	2	0	25	2	0	7	2	0	1	96
3	13	0	3	14	1	3	0	4	3	11	0	142

图 4　特征向量部分结果示意图

1.3　分类器训练与测试

得到各视频样本的特征向量后,进入动作识别分类器的训练与测试阶段。目前,比较常用的分类器有 Adaboost、K-近邻(K-NN)和 SVM。考虑到 Adaboost 用于多分类时具有过分依赖弱分类器的不足而 K-NN 具有计算量大和分类结果受 K 值影响较大的缺点,所以本文采用 SVM 实现动作识别分类器的训练和测试。SVM 处理分类问题的基本思想是:通过定义一个线性最优超平面从而将分类问题转化为确定该超平面的凸优化问题。

训练动作识别分类器时,需将经过量化处理的特征向量作为非线性 SVM 的输入。本文的动作识别属于多分类问题,由于当采用“一对多”的方式解决多分类问题向二分类问题的转化时,具有优化问题的规模小、分类速度快的优点,所以本文采用“一对多”的方式训练分类器模型。同时,又由于 SVM 中核函数的选取对分类结果的影响至关重要,所以,有别于通常使用的 χ^2 核函数,本文选择直方图交叉核函数[17]作为训练 SVM 分类器时的核函数,后续实验结果也验证了该核函数的优越性。直方图交叉核函数具体的表达式如式

$$K(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sum_{n=1}^{m}\min\{a_n,b_n\} \tag{3}$$

所示。其中 $\boldsymbol{X}_i$、$\boldsymbol{X}_j$ 为两个任意的特征向量,a_n、b_n 分别为 $\boldsymbol{X}_i$、$\boldsymbol{X}_j$ 第 n 维的特征值,m 为特征向量的维度。由式(3)可以看出,直方图交叉核函数具有计算复杂性低的特点,这有助于 SVM 将低维空间中线性不可分的问题快速转化为高维空间中线性可分的问题。图 5 为根据后文 KTH 数据集中 384 个训练样本的特征向量绘制的直方图交叉核函数的示意图。其中,X、Y 轴分别代表 i 和 j,即特征向量的编号,Z 轴代表所得函数值。由图 5 可以看出,当 $\boldsymbol{X}_i = \boldsymbol{X}_j$ 时,直方图交叉核函数在 XY 平面的对角线上取得最大值,且此最大值为 $\boldsymbol{X}_i$ 自身各列的和。由此可知,当两个样本越是相似时,得到的函数值越大;相

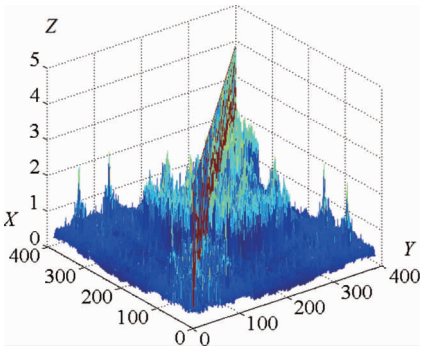


图 5　直方图交叉核函数的示意图

反,当两个样本越是不同时,得到的函数值就越小,从而有利于 SVM 区分不同动作。动作识别分类器训练完成后,即可将测试样本的特征向量输入该分类器模型进行动作识别并得到动作识别结果。

1.4 动作识别算法

算法名称:ActionRecognitionAlgorithm;输入:动作视频;输出:动作识别结果。

步骤 1:跟踪视频中运动物体上的时空特征点形成运动轨迹。

步骤 2:分别计算轨迹的 HOG、HOF、MBH 及 LBP 特征,然后进行联合形成联合特征。

步骤 3:对训练样本特征矩阵平均采样进行 K-means 聚类构造出由定量单词组成的码书。

步骤 4:利用码书量化所有样本特征矩阵得到新的特征向量。

步骤 5:将训练样本的特征向量输入支持向量机训练动作识别分类器。

步骤 6:将测试样本的特征向量输入已训练好的分类器进行动作识别。

2 实验与分析

2.1 数据集及实验方案

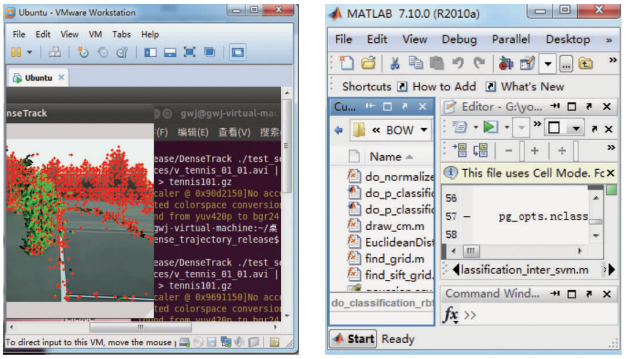
动作识别领域中较为常见的数据集主要有 KTH[18]、Weizmann[19]、UCF Sports[20] 和 YouTube[21] 等。由于 KTH 数据集具有较强的代表性而受到广泛使用,同时由于 YouTube 数据集具有较大的挑战性,所以本文选用 KTH 和 YouTube 数据集进行了实验。实验硬件环境为 CPU Intel Xeon 六核 2.0GHz,内存 16G。特征提取功能采用 C + + 语言在 Linux 下编程实现,特征处理及分类器的训练与测试功能采用 Matlab 语言在 Win 7 下编程实现。图 6 展示了两个数据集中每种动作的样本帧。其中,第一行为从 KTH 数据集中截取的样本帧,实际视频帧大小为 160 × 120 ,帧频为 25 帧/秒;第二、三行为从 YouTube数据集中截取的样本帧,实际视频帧大小为 320 × 240 ,帧频为 29 帧/秒。

KTH 数据集中共包含 6 种人体动作:拳击、拍手、慢跑、快跑、散步和挥手。每种动作分别由 25 个人在四种场景下录制而成。此数据集具有背景不变的特点。为了使实验结果具有可比性,实验过程中,我们根据制作者最初的设置,将样本集分为包括第 2、3、5、6、7、8、9、10、22 个人的测试样本集和另外 16 个人的训练样本集,并利用所有训练样本进行一次分类器模型的训练。

YouTube 数据集中共包含 11 种人体动作:秋千、网球、蹦床、排球、散步、潜水、投篮、骑自行车、打高尔夫、骑马和颠球。由于像机运动、视角变换、背景混乱及光照条件变化等原因使得此数据集的识别十分富有挑战性。同样,实验过程中我们按照原制作者的实验方案利用此数据集,即采用留一法交叉验证,分组进行 25 次分类器模型的训练,以保证每个样本都参与一次测试。

图 6 实验所用数据集中的样本帧

图7展示了动作识别的操作界面。其中,图7(a)所示为特征提取界面,图7(b)所示为特征处理及分类器的训练与测试界面。



(a) (b)

图7 动作识别操作界面图

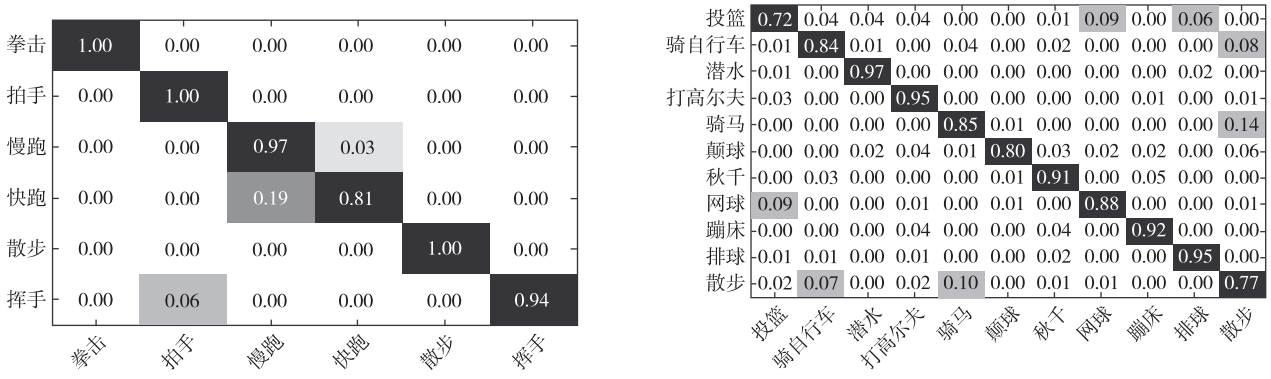
2.2 实验结果及分析

按照上述实验方案,得到两个数据集上的动作识别结果及其混淆矩阵。图8以KTH数据集中第2个人在场景1下的动作“拳击”,“拍手”和“快跑”为例,通过分别截取三个视频样本的第26帧、第51帧、第251帧和最后一帧展示了动作识别结果。其中,图8(a)和8(b)为正确识别的结果,图8(c)为将动作“快跑”误识为动作“慢跑”的结果。两个数据集上识别结果的混淆矩阵如图9所示。

由图9可知,在KTH数据集识别结果的混淆矩阵(a)中,大部分混淆发生在动作“快跑”与“慢跑”之间,这是由于这两种动作的各种特征都比较类似,唯一不同的是它们的速度。同样,在YouTube数据集识别结果的混淆矩阵(b)中,由于动作“投篮”与

(a)拳击 (b)拍手 (c)快跑

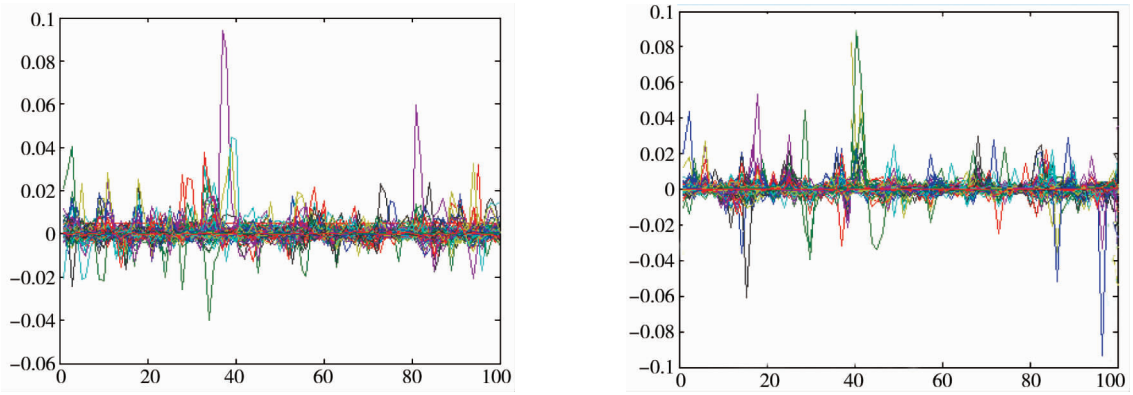
图8 部分动作识别结果



(a) KTH数据集识别结果的混淆矩阵 (b) YouTube数据集识别结果的混淆矩阵

图9 两个数据集识别结果的混淆矩阵

动作“网球”和“排球”引发的光流场以及运动轨迹的梯度方向基本一致,使得动作“投篮”容易被误识。具体分析如下:随机提取动作“投篮”与动作“排球”各100个样本的特征向量,分别截取特征向量中代表HOG、HOF部分的特征值,然后分别对两种动作的特征向量中相应部分的特征值做减法,将得到的差值分别投影到二维平面得到图10所示的特征差值图。由图10可以看出,两种动作在HOG和HOF特征上的差值较小。类似地,动作“投篮”与动作“排球”也有相似的结果。由此可知,动作“投篮”容易被误识。此处需说明的是:容易被误识是指动作“投篮”的识别率只是相对较低,大多数情况下还是能够正确识别的(实验中的识别率为72%),因为还需要综合考虑MBH和LBP两种特征的实际情况。



(a) HOG 特征的差值图　　(b) HOF 特征的差值图

图 10　动作“投篮”与动作“网球”的特征差值图

同时，为验证本文方法的有效性，我们分别按照以下四种方法进行了实验：(1) HOG、HOF、MBH 三种特征联合且选用 χ^2 核函数的动作识别方法；(2) HOG、HOF、MBH 三种特征联合且选用直方图交叉核函数的动作识别方法；(3) LBP、HOG、HOF、MBH 四种特征联合且选用 χ^2 核函数的动作识别方法；(4) LBP、HOG、HOF、MBH 四种特征联合且选用直方图交叉核函数的动作识别方法。四种方法在 KTH 和 YouTube 两个数据集上的识别率如表 1 所示。其中，识别率 $=\frac{\text{正确识别的样本数}}{\text{参与测试的总样本数}}\times 100\%$。

表 1　四种实验方法在两个数据集上的识别率

数据集	无 LBP + χ^2	无 LBP + 直	有 LBP + χ^2	有 LBP + 直
KTH	94.0%	94.4%	94.9%	95.4%
YouTube	84.2%	85.0%	86.1%	86.9%

表 1 的实验结果表明：(1) 在选用相同核函数的情况下（无论是 χ^2 核函数还是直方图交叉核函数），有 LBP 特征的动作识别效果皆优于无 LBP 特征的动作识别效果，由此验证了 LBP 特征对动作识别的有效性；(2) 不管在有无 LBP 特征的情况下，选用直方图交叉核函数的动作识别效果均优于选用 χ^2 核函数的动作识别效果，由此验证了本文方法中选用直方图交叉核函数进行动作识别的有效性。

为了进一步评价本文方法的识别效果，将其与目前识别效果较好的其他方法分别在 KTH 和 YouTube 两个数据集上进行了对比实验，所得各种动作的识别结果分别如表 2 和表 3 所示。

表 2　本文方法与现有方法在 KTH 数据集上识别率的比较

动作类别	识别方法			
	Gilbert[22] 方法	Liu[23] 方法	Ji[24] 方法	本文方法
拳击	100.0%	96.0%	95.2%	100.0%
拍手	94.0%	95.0%	94.1%	100.0%
慢跑	99.0%	83.8%	83.0%	97.0%
快跑	91.0%	85.4%	96.5%	81.0%
散步	89.0%	90.7%	96.2%	100.0%
挥手	94.0%	98.7%	93.4%	94.0%
平均识别率	94.5%	91.6%	93.1%	95.4%

表 3　本文方法与现有方法在 YouTube 数据集上识别率的比较

动作类别	识别方法			
	Wang[13] 方法	Liu[21] 方法	Zhang[25] 方法	本文方法
投篮	43.0%	53.0%	98.0%	72.0%
骑自行车	91.7%	73.0%	74.0%	84.0%
潜水	99.0%	81.0%	80.0%	97.0%
打高尔夫	97.0%	86.0%	68.0%	95.0%
骑马	85.0%	72.0%	65.0%	85.1%
颠球	76.0%	54.0%	67.0%	80.0%
秋千	88.0%	57.0%	71.0%	91.0%
网球	71.0%	80.0%	68.0%	88.0%
蹦床	94.0%	79.0%	80.0%	92.0%
排球	95.0%	73.3%	77.0%	95.0%
散步	87.0%	75.0%	54.0%	77.0%
平均识别率	84.2%	71.2%	72.9%	86.9%

由表2可知,本文方法对动作“拳击”、“拍手”、“散步”的识别率明显优于其他方法,同时在动作“慢跑”和“挥手”上也获得了较好的识别结果,且总体识别率比其他方法都高。这是由于本文不仅关注了运动本身的特征,还将运动轨迹的特征与所在视频帧局部区域的纹理信息进行了结合,从而有效地区分了每种动作的轨迹形状、运动特征和主要活动区域。例如,在其他方法中容易被混淆的“拍手”与“挥手”动作,在本文中却取得了较好的识别效果。这是因为虽然由动作“拍手”和“挥手”产生的轨迹形状特征和引发的光流场比较类似,但是由于“拍手”动作主要发生在人体胸前部位,而“挥手”的活动区域主要位于人体两侧和头顶上方,所以本文方法通过计算运动轨迹所在视频帧局部区域的纹理信息即可很好地对这两种动作进行区分。遗憾的是,本文方法对动作“快跑”的识别率低于其他几种方法。我们认为主要原因是由于从动作“快跑”视频中提取到的运动轨迹一部分由人体上半身引起,这部分的轨迹特征与动作 Jogging 相似性较大,所以干扰了识别结果。

此外,我们分析其他方法整体识别率低的主要原因在于:对于对动作“慢跑”和“挥手”的识别结果分别有明显优势的 Gilbert[22] 方法和 Liu[23] 方法来说,仅利用2维角点的简单尺寸、通道信息和动作属性特征来表示人体运动并不能涵盖全面的运动特征,所以导致 Gilbert[22] 方法和 Liu[23] 方法的整体识别水平有所下降。对 Ji[24] 方法而言,由于其通过分别计算 xt 与 yt 两个平面上时空特征点的梯度后合并成为特征点的 CHOG3D 特征,使得计算得到的“慢跑”动作的特征与“快跑”和“散步”两种动作的特征皆比较相似,导致“慢跑”动作容易被误识为“快跑”和“散步”动作,从而降低了“慢跑”动作的识别率,进而也降低了整体识别率。

由表3可知,本文方法对“骑马”、“颠球”、“秋千”、“网球”和“排球”5种动作的识别率优于其他几种方法(除动作“排球”与 Wang[13] 方法相等外),而且在动作“骑自行车”、“潜水”、“打高尔夫”、“蹦床”和“散步”上也获得了可观的识别结果。在对上述动作的识别过程中,当摄像机移动引发视频图像光照条件改变时,由于本文引入的 LBP 特征具有光照不变性,从而克服了由光照条件改变对动作识别带来的不利影响。而且,在选用直方图交叉核函数进行分类器的训练与测试时,采用了对两样本对应维度上较小的值求和并将结果映射到高维空间的处理方式,从而过滤掉了一部分由摄像机移动产生的轨迹特征。所以,本文方法在总体识别率上要优于其他方法。对其他方法而言,Wang[13] 方法由于缺少对运动主体与整体视频帧之间相对关系的分析,使得动作“投篮”的识别率偏低,影响了整体的识别效果。Liu[21] 方法通过结合视频中的静止表观特征与局部运动特征进行动作识别,忽略了由摄像机移动产生的噪声的影响,使得该方法对动作“投篮”、“颠球”和“秋千”的识别结果并不理想,如表3所示。而在 Zhang[25] 方法中,由于其对多人行为的交互进行了有针对性的分析,所以对动作“投篮”的识别率较高,但由于没有对提取到的运动信息做进一步处理,导致将摄像机运动与人体运动产生的运动信息相混淆,使得该方法对“打高尔夫”、“骑马”和“散步”等明显有摄像机移动的室外动作的识别率偏低。所以,总体比较而言,本文方法获得了较好的识别结果。

为了进一步衡量本文所提动作识别方法的执行效率,在采用相同特征提取方法和量化处理方式及相同实验环境下,我们分别用直方图交叉核函数与普遍采用的 χ^2 核函数作为非线性 SVM 的核函数训练动作识别分类器并进行测试。为使比较结果更加直观,我们仅对与两种核函数有关的训练时间和测试时间进行比较。训练时间计时从向 SVM 输入所有训练样本特征向量开始到输出分类器模型为止,测试时间计时从输入所有测试样本特征向量开始到输出动作识别结果为止。实验中,我们利用 Matlab 中的 tic 命令作为计时开始标记,toc 命令作为计时结束标记,分别得到了训练和测试过程的时间消耗。具体时间对比结果如表4所示。

表4　两种核函数在相同条件下所用时间的比较

数据集	核函数	训练时间(s)	测试时间(s)
KTH	χ^2	155.71	93.48
	直方图交叉	5.78	3.16
YouTube	χ^2	1551.90	79.44
	直方图交叉	80.53	4.21

由表4可知,采用 χ^2 核函数进行训练与测试时,由于涉及到特征通道问题,所以需要分别计算每个特征向量在每种特征通道上的 χ^2 距离,从而使得耗时较长。而采用直方图交叉核函数作为动作识别分类器的核函数时,仅需要比较两个特征向量对应

维度上值的大小并取较小值的和映射到高维空间即可,所以有利于 SVM 快速寻找到最优的分类超平面,从而明显缩短了训练与测试时间。由此可见,由于两个核函数对训练样本与测试样本特征向量的处理方式不同,使得它们对应的训练时间和测试时间也不同。

3 结 论

为了更好地识别出视频序列中的人体动作,本文提出了一种新颖的基于特征联合和直方图交叉核函数的动作识别方法。所提方法主要有以下三点贡献:(1)在已有轨迹特征的基础上,通过加入 LBP 特征,将运动轨迹信息与所在视频帧局部的纹理信息相结合,提高了动作识别的准确率。(2)利用 bag-of-features 方法进行特征处理时,提出平均采样构造码书的方法,解决了识别结果不易重现的问题。(3)采用直方图交叉核函数作为 SVM 进行动作分类时的核函数,不仅进一步提高了动作识别的准确率,而且缩短了识别时间。实验结果验证了所提方法的可行性和有效性。

参考文献

[1] Bobick A F, Davis J W. The Recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23(3): 257-267

[2] 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. 自动化学报,2012,38(8):1380-1384

[3] Weinland D, Boyer E. Action recognition using exemplar-based embedding. In: Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008. 3033-3039

[4] Yoon S M, Kuijper A. Human action recognition using segmented skeletal features. In: Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010. 3740-3743

[5] Marszalek M, Laptev I, Schmid C. Actions in context. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009. 2929-2936

[6] Ikizler-Cinbis N, Sclaroff S. Object, scene and actions: combining multiple features for human action recognition. In: Proceedings of the 11th European Conference on Computer Vision, Crete, Greece, 2010. 494-507

[7] 王传旭,刘云,厉万庆. 基于时空特征点的非监督姿态建模和行为识别的算法研究. 电子学报,2011,39(8):1751-1756

[8] Matikainen P, Hebert M, Sukthankar R. Representing pairwise spatial and temporal relations for action recognition. In: Proceedings of the 11th European Conference on Computer Vision, Crete, Greece, 2010. 508-521

[9] Ma Z, Yang Y, Hauptmann A G, et al. Classifier-specific intermediate representation for multimedia tasks. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 2012. 50:1-8

[10] Scovanner P, Ali S, Shah M. A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the ACM International Multimedia Conference and Exhibition, Augsburg, Germany, 2007. 357-360

[11] Klaser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of British Machine Vision Conference, Leeds, UK, 2008. 995-1004

[12] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 2008, 110(3): 346-359

[13] Wang H, Klaser A, Schmid C, et al. Action recognition by dense trajectories. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, 2011. 3169-3176

[14] Li F, Pietro P. A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005. 524-531

[15] Gunnar F. Two-frame motion estimation based on polynomial expansion. In: Proceedings of the 13th Scandinavian Conference on Image Analysis, Halmstad, Sweden, 2003. 363-370

[16] Wang H, Ullah M M, Klaser A, et al. Evaluation of local spatio-temporal features for action recognition. In: Proceedings of British Machine Vision Conference, London, UK, 2009. 1-11

[17] Barla A, Odone F, Verri A. Histogram intersection kernel for image classification. In: Proceedings of the IEEE International Conference on Image Processing, Barcelona, Spain, 2003. 513-516

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

[19] Gorelick L, Blank M, Shechtman E, et al. Action as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(12): 2247-2253

[20] Rodriguez M D, Ahmed J, Shah M. Action MACH: a spa-

tio-temporal maximum average correlation height filter for action recognition. In:Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008. 3001-3008

[21] Liu J, Luo J, Shah M. Recognizing realistic actions from videos "in the wild". In:Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009. 1996-2003

[22] Gilbert A, Illingworth J, Bowden R. Fast realistic multi-action recognition using mined dense spatio-temporal features. In:Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 2009. 925-931

[23] Liu J, Kuipers B, Savarese S. Recognizing human actions by attributes. In:Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, 2011. 3337-3344

[24] Ji Y, Shimada A, Nagahara H, et al. A compact descriptor CHOG3D and its application in human action recognition. *IEE J Transactions on Electrical and Electronic Engineering*, 2013, 8(1):69-77

[25] Zhang Y, Liu X, Chang M, et al. Spatio-temporal phrases for activity recognition. In:Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 2012. 707-721

An action recognition approach based on feature combination and histogram intersection kernel function

Zhang Shihui[*,**], Gao Wenjing[*], Kong Lingfu[*,**]

([*] College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

([**] The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004)

Abstract

In order to improve the recognition rate and real-time performance of action recognition, a novel action recognition approach based on the feature combination and histogram intersection kernel function is proposed. The approach tracks the local spatio-temporal feature points on the moving object in a video sequence to form motion trajectories, and calculates each trajectory's HOG feature, HOF feature, MBH feature and the LBP feature of each point on the trajectory to form a joint feature matrix, and then, equivalently samples each action's joint feature matrices of each training sample and uses the bag-of-features method to do the K-means clustering to form a codebook after the combination of the sampling results. On this basis, the joint feature matrix of each sample is quantitated by using the codebook to obtain the eigenvector which represents the motion information and the structure information in the video sample. Finally, the action recognition classifier is trained and tested by taking the eigenvector as the input of a SVM and selecting the histogram intersection kernel function as the kernel function of the SVM. The experimental results show that the proposed approach not only can improve the recognition rate, but also can shorten the training and testing time by using the histogram intersection kernel function.

Key words: action recognition, motion trajectory, joint features, bag-of-features, histogram intersection kernel function