doi:10.3772/j.issn.1002-0470.2015.08-09.005

基于随机投影和 Fisher 向量的人的行为识别①

何 军② 薛 莹 胡昭华 孙 伟

(南京信息工程大学电子与信息工程学院 江苏省气象探测与信息处理重点实验室 南京 210044)

摘 要 为了提高识别视频人物行为的效果,在定义和分析高斯混合模型(GMM)的基础上,提出了一种基于随机投影(RP)和 Fisher 向量(FV)的行为识别方法。该方法通过随机投影将高维轨迹描述子投影到低维子空间来实现特征轨迹的降维,然后利用 GMM-FV 混合模型对降维后的轨迹特征向量进行空间聚类编码,以提高行为识别的准确率,最后再利用随机投影对 Fisher 编码向量进行二次降维以降低计算复杂度。用 KTH 和 UCF50 两种数据集进行的试验表明,与现有跟踪识别算法相比,该方法降低了计算的复杂度,提高了行为识别的准确率,在两种数据集上的识别都表现出了良好的鲁棒性。

关键词 行为识别, 特征轨迹, 随机投影(RP), 特征降维, Fisher 向量(FV)

0 引 言

随着各种社交网络、视频点播工具和多媒体设备在社会生活中的不断普及,以及各种在线视频共享技术的日趋成熟,针对视频的行为识别技术已被广泛运用到视频监控、视频检索、视频注释、行为军事检测、医疗诊断和监护等领域,具有广阔的应用前景和经济价值[1-3]。因此对视频人物行为的识别技术的探索研究就显得尤为重要。

简要地说,视频行为识别过程包括三个步骤[4]:(1)对一视频行为提取局部特征描述子,如尺度不变特征变换(scale invariant feature transform, SIFT)[5],方向梯度直方图(histogram of oriented gradient, HOG)[6],光流直方图(histogram of optical flow, HOF)[7]等;(2)将这些轨迹特征编码成固定长度的高维向量作为整个视频行为的全局表达向量;(3)利用训练好的分类器实现人的行为的分类。文献[8]将视频在时间域上切分为若干个网格,通过计算直方图特征,再利用主成分分析(PCA)对特征降维,最后将其映射到一个视觉词袋(bag-of-words,BoW)模型[9]中以实现目标分类。文献[10]利用稀疏表示视频动作的时空特征,然后通过学习完备的字典实现对人的行为的建模。文献[11]通过稀疏表示步态能量图和运动描述子这两种特征,实现了人的手势的识别。文献[12]利用随机投影定理实现对高维数据的降维,并将降维后的特征用于图像和文本数据的分类。本研究从前人工作中得到了启发,通过随机投影(random projection,RP)[13]的方法将高维轨迹描述子投影到一个低维子空间实现特征轨迹的降维,再利用高斯混合模型(Gaussion Mixture Model, GMM)混合费舍尔向量(Fisher vector, FV)的模型(简称 GMM-FV 混合模型)[14]对降维后的轨迹特征向量进行空间聚类编码。这样就从统计学的意义上描述了这些特征轨迹描述子的概率分布情况,不仅丰富了行为动作的特征表达,还降低了分类器训练的复杂度。最后再次利用随机投影的方法对 Fisher 编码向量进行二次降维,试验表明该方法能有效地提高运行效率和行为识别的准确率。

① 国家自然科学基金(NSFC61203273, NSFC61304205),江苏省自然科学基金(BK20141004)和大学生创新创业训练计划(201310300010Z)资助项目。

② 男,1978 年生,博士,副教授;研究方向:大数据机器学习,计算机视觉等;联系人,E-mail: jhe@nuist.edu.cn

(收稿日期:2015-05-14)

1 基于随机投影和 Fisher 向量的行为特征

1.1 特征轨迹提取

本文通过提取稠密的轨迹[15]来表征一类行为运动,利用光流场对兴趣点进行多层次稠密采样来实现跟踪。这些兴趣点沿一个密集的网格被重复采样且被跟踪在一个固定长度帧的范围内,行为轨迹就是这些固定帧数内的特征描述子连续性表达的结果。轨迹的形状用来区别不同的人物行为变化,它表现在视频中就是人物目标在视频中的运动位置在时间和空间上的改变,即位移矢量。

在图像和视频处理过程中,得到的特征向量一般都是高维的。考虑到人的行为在每个视频中出现的位置的不同,本文通过对提取到的所有位置信息进行求和运算,实现位置矢量的归一化。对于任意一条轨迹来说,除了提取其位置信息外,还要通过在时间和空间变化过程中产生的各个描述子信息来丰富它的表达。如方向梯度直方图(HOG)用来描述人的外在的静态信息,光流直方图(HOF)用来描述轨迹的局部运动信息,而运动边界直方图(motion boundary histogram, MBH)[16]用来描述像素之间的相对运动。因此,我们最终确定的轨迹是位置信息、方向梯度、光流和运动边界直方图信息的集合。文献[15]中介绍了关于 HOG、HOF、MBH 的描述子维度信息的计算方法,本文就不再重复叙述。于是在固定长度帧($L = 15$)的视频流中,轨迹信息的描述子的维度包括 30 维的位置信息(x 轴和 y 轴)、480 维的方向梯度信息、540 维的光流信息和 960 维的运动边界信息(x 方向和 y 方向),将这些描述子信息级联起来,则视频每 15 帧中各个被跟踪的兴趣点产生的轨迹长度就是一个 2010 维的特征向量。用数学语言进行如下简单的描述:设 $\boldsymbol{X}_I = [x_1, \cdots, x_{T_I}]^{\mathrm{T}}$ 表示第 I 个行为视频的轨迹特征向量,其中 x_{T_I} 表示级联后各个局部描述子信息, T_I 表示第 I 个视频中提取到有效轨迹的数目,那么由 m 个训练视频的特征轨迹向量所组成的行为轨迹矩阵就可以表示成 $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m]^{\mathrm{T}}$, 该矩阵就是我们最终需要的行为特征轨迹信息。

1.2 特征降维

1.2.1 随机投影

一个完整的行为视频是由许多个这样的高维特征轨迹向量来表示的,为了提高分类器训练的准确率,需要大量的行为视频来构造这样的轨迹矩阵。面对如此庞大的高维矩阵,我们不得不采取措施来降低这些轨迹特征向量的维度,这样既能减小计算复杂度还能保留编码特征的原始信息。在随机投影理论[17]中,对于含有 N 条 D 维特征描述子轨迹的特征空间 $x_t \in \boldsymbol{R}^D$, 对其作用一个列单元长度的随机矩阵 $\boldsymbol{\Phi}$,将其投影到一个低维子空间 $v_t \in \boldsymbol{R}^d$ 中,其中 $d << D$, 其公式表达如下:

$$v_t^d = \boldsymbol{\Phi} x_t^D \tag{1}$$

Johnson-Lindenstrauss(JL)引理[18]表明,将数据的高维特征空间映射到低维特征空间后数据点之间距离基本保持不变。因此如果随机矩阵 $\boldsymbol{\Phi}$ 满足 JL 引理,就可以将 $x_t \in \boldsymbol{R}^D$ 以最小误差从 $v_t \in \boldsymbol{R}^d$ 重构出来,即该定理保证了投影后的低维子空间 v_t 基本包含了原始信号 x_t 中的全部信息,这对利用低维子空间来分析高维数据信号来说是一个强有力的理论支持。JL 引理高效快速的特性使其在其他算法中也得到了很好的应用,文献[19]提出了一种快速的实时压缩跟踪算法,该算法跟踪速度能够达到 40 帧/秒且跟踪鲁棒性较好。

1.2.2 随机投影矩阵的构造

Candes 和 Tao 在文献[20]中提出了约束等距性(restricted isometry property, RIP)准则的概念,它用来描述测量矩阵 $\boldsymbol{\Phi}$ 所满足的约束条件:如果所有满足 $\| x \|_0 \leqslant \boldsymbol{K}$ 的 x 为满足

$$(1 - \delta_k) \| x \|_2^2 \leqslant \| \boldsymbol{\Phi} x \|_2^2 \leqslant (1 + \delta_k) \| x \|_2^2 \tag{2}$$

的最小值,那么称 $\boldsymbol{\Phi}$ 以参数 (K, δ_k) 满足 RIP 线性算子。典型的满足 RIP 准则的测量矩阵就是随机高斯矩阵 $\boldsymbol{\Phi} \in \boldsymbol{R}^{n \times m}$, 其中矩阵元素 $\phi_{i,j}$ 满足 $N(0,1)$ 分布。但是当 m 的维数比较大时,这个矩阵仍然是比较稠密的,其运算和存储消耗还是相当的大。因此可以定义一个非常稀疏随机测量矩阵:

$$\phi_{i,j} = \sqrt{s} \times \begin{cases} 1, & p = \dfrac{1}{2s} \\ 0, & p = 1 - \dfrac{1}{s} \\ -1, & p = \dfrac{1}{2s} \end{cases} \tag{3}$$

Achlioptas[18]证明,当 $s = 2$ 或 3 时,测量矩阵是一个稀疏矩阵且满足 JL 引理,这样就减少 2/3 的计算量。甚至当 $s = m/\log(m)$ 时,该测量矩阵进行随机投影的精度和高斯随机矩阵的投影精度达到了一致性,且对于每个测量矩阵只需要计算每列非零个数,这样大大简化了计算量,为保证测量矩阵的稀疏性,本文取 $s = m/4$。假设 c 表示测量矩阵 $\boldsymbol{\Phi}_{d\times N}$ 每列非零项个数,N 表示 N 条 d 维特征描述子轨迹,比较随机投影(RP)、主成分分析(PCA)和奇异值分解(SVD)这三种方法的计算复杂度可知,RP 的计算复杂度最低,如表 1 所示。

表 1　随机投影、主成分分析法和奇异值分解法的计算复杂度

	RP	SVD	PCA
计算复杂度	$O(ckN)$	$O(cdN)$	$O(d^2N) + O(d^3)$

虽然在均方意义上效果最好的是用 PCA 来实现数据的降维,但对高维数据集来说其计算过程是相当耗时的。而随机投影理论是将原始的高维数据随机地投影到一个低维子空间中,计算速度快且不会引入不必要的数据失真,因此本文引入随机投影的方法来实现轨迹特征的降维。

1.3　GMM-FV 混合模型

词袋(BoW)模型是最常用于图像视频分类的直方图模型,它通过提取视频中的局部轨迹特征来构建一个丰富的视觉词典,并利用中心聚类的方式分别统计出局部特征向量相对于中心单词出现的频率,由这些视觉词频构成的直方图表示一类行为视频,最终达到人的行为识别的目的。BoW 模型最关键的就是要构造出一个非常庞大的视觉词典,因而行为识别的准确率在很大程度上取决于所构造特征词典的规模的大小,这在一定程度上就增加了计算成本和时间消耗。

本研究采用 GMM-FV 混合模型。不同于 BoW 模型的硬划分,GMM-FV 混合模型对轨迹特征向量进行软划分,它主要结合空间聚合(pooling)的思想,在性能和识别准确率上相比于 BoW 模型都有了很大的提高。Fisher 向量[21]融合了 Fisher 核生成模式和判别模式的特点,不仅能计算出每个特性描述子出现的频率,还能从统计学的意义上描述这些特征描述子的概率分布情况,既丰富了行为动作的特征表达又提高了行为识别的效率。

在计算机视觉领域中,图像中局部特征的生成过程所采用的高斯混合模型(GMM)[22]可以被看作是一个通用的视觉模型,可以用来逼近任何连续的概率函数分布。设 p_λ 是概率密度函数,用来对特征轨迹信号的生成过程建模,其中 $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_M]$ 表示 p_λ 的 M 个参数向量。$\boldsymbol{X}_I = [x_t, t = 1, \cdots, T_I]$ 表示从一个行为视频中提取到有效轨迹特征的集合,T_I 表示该行为视频中轨迹的数目。假设每个局部特征 $x_t \in \boldsymbol{R}^d$ 都服从独立同分布,d 表示轨迹特征经过降维处理后的维度,那么对含有 K 个高斯单元 GMM 的参数集 $\lambda = \{w_i, u_i, \sum_i\}$, $i = 1, \cdots, K$,其高斯混合模型的定义如下:

$$p_\lambda(x_t) = \sum_{i=1}^{K} w_i p_i(x_t) \tag{4}$$

其中 w_i、u_i 和 $\sum_i$ 分别表示第 i 个高斯单元的混合权重、均值向量和协方差矩阵,$p_i(x_t)$ 表示特征轨迹向量 x_t 的第 i 个高斯单元,表达式为

$$p_i(x_t) = \frac{\exp\{-\frac{1}{2}(x_t - u_i)^T \sum_i^{-1} (x_t - u_i)\}}{(2\pi)^{d/2} |\sum_i|^{1/2}} \tag{5}$$

假定协方差矩阵为对角矩阵,由贝叶斯公式可知,轨迹特征 x_t 分配到第 i 个高斯单元的概率定义为

$$r_t(i) = \frac{w_i p_i(x_t)}{\sum_{k=1}^{K} w_k p_k(x_t)} \tag{6}$$

设 $\ell_\lambda(\boldsymbol{X}) = \log p_\lambda(\boldsymbol{X}) = \sum_{t=1}^{T} \log p_\lambda(x_t)$ 是 $\boldsymbol{X}$ 关于 λ 的对数似然函数,则局部特征轨迹 x_t 关于 GMM 参数集 $\lambda = \{w_i, u_i, \sum_i\}$, $i = 1, \cdots, K$ 的梯度分别表示为

$$\frac{\partial \mathcal{L}_\lambda(\boldsymbol{X})}{\partial w_i} = \sum_{t=1}^{T}\left[\frac{r_t(i)}{w_i} - \frac{r_t(1)}{w_1}\right] \tag{7}$$

$$\frac{\partial \mathcal{L}_\lambda(\boldsymbol{X})}{\partial u_i^k} = \sum_{t=1}^{T} r_t(i)\left[\frac{x_t^k - u_i^k}{(\sigma_i^k)^2}\right] \tag{8}$$

$$\frac{\partial \mathcal{L}_\lambda(\boldsymbol{X})}{\partial \sigma_i^k} = \sum_{t=1}^{T} r_t(i)\left[\frac{(x_t^k - u_i^k)^2}{(\sigma_i^k)^3} - \frac{1}{\sigma_i^k}\right] \tag{9}$$

其中，$(\sum_i) = ((\sigma_i^1)^2, \cdots, (\sigma_i^K)^2)$，$\sigma_i^k$ 表示协方差矩阵 $\sum_i$ 中的标准方差。Fisher 向量就是由这些相对于平均值和标准偏差等参数的偏导数之间的级联而成的。考虑到 d 表示的是轨迹特征经过降维后的维度，K 表示该模型中高斯单元的个数，那么得到的 Fisher 向量的维度就是 $2dK$。由于最终的 Fisher 向量也是稀疏和高维的，因此利用 1.2 节提到的随机投影理论，可以实现对编码轨迹进行二次投影降维。试验表明二次降维后的特征向量大大节约了训练分类器的时间，提高了行为识别的识别率。

2　行为识别算法

我们针对一类行为视频，首先在固定帧数的前提下提取和跟踪局部行为特征，再在最小误差允许范围内提取有效轨迹，然后融合各类描述子信息形成一个高维轨迹特征向量，组成该类行为视频的特征轨迹矩阵空间，最后嵌入到 GMM-FV 混合模型框架中，用一个支持向量机（SVM）分类器通过添加类别标签的方式训练出用于区别各种行为的一个超平面，通过这个超平面实现最终的行为分类。图 1 所示为本文提出的基于 Fisher 向量和投影定理的行为识别的整个流程。期间我们采用随机投影的方式对高维特征进行二次投影降维来降低计算复杂度，具体算法见算法 1。

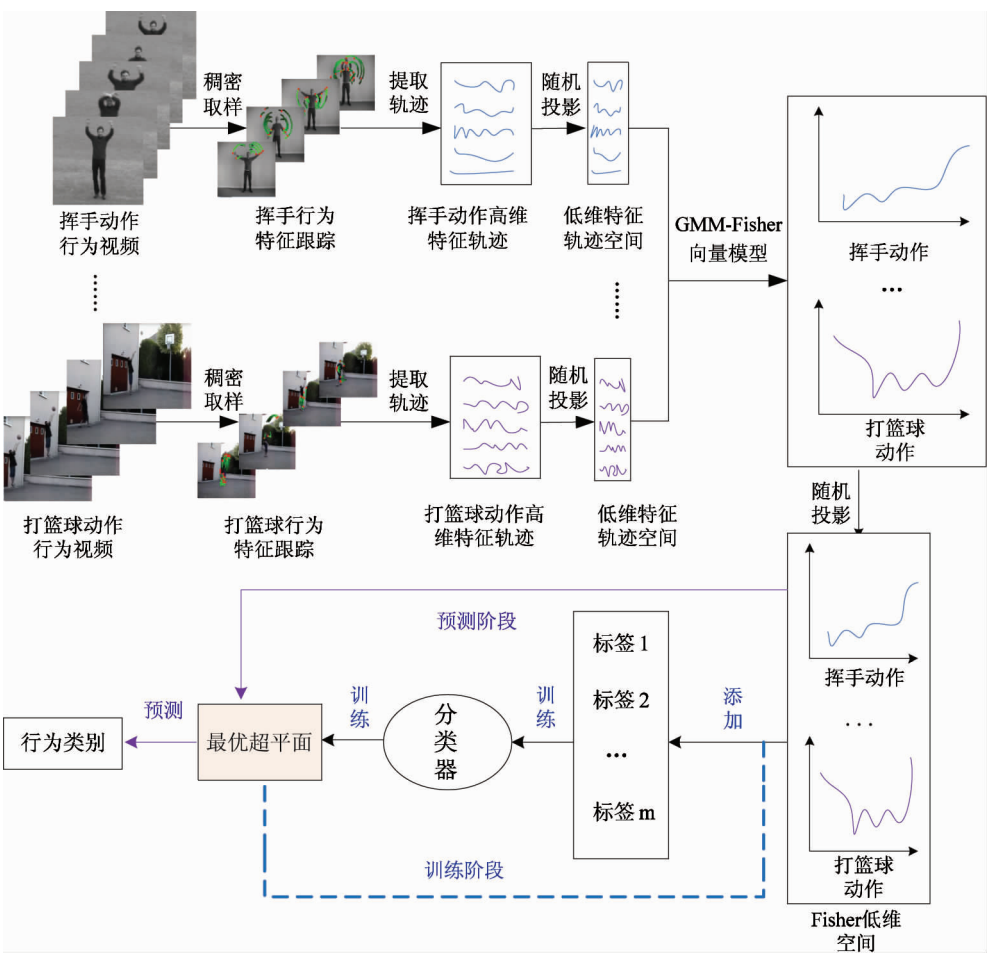

图 1　基于随机投影和 Fisher 向量的行为识别过程

算法 1. 基于 Fisher 向量和随机投影的行为识别算法

训练部分:

输入:训练视频 $S = [S_1, S_2, \cdots, S_m]$, 标签 $L = [l_1, l_2, \cdots, l_m]$, m 表示训练视频个数。

输出:可用于行为识别的分类超平面。

预测部分:

输入:测试视频 $Z = [Z_1, Z_2, \cdots, Z_n]$, n 表示测试视频个数。

输出:行为识别的类别标签。

开始

训练部分:

1 对一个训练视频 S_I 提取特征轨迹描述子形成向量 $\boldsymbol{X}_I = [x_1, \cdots, x_{T_I}]^{\mathrm{T}}$。

2 遍历所有视频 $S = [S_1, S_2, \cdots, S_m]$ 组成高维特征轨迹矩阵 $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_m]$。

3 将高维的轨迹特征向量随机投影到一个低维子空间 $\boldsymbol{V}^{\mathrm{RP}} = [v_t \in \boldsymbol{R}^d]$, $t = 1, \cdots, \sum_{I=1}^{m} T_I$ 中,其中, d 表示轨迹特征经随机降维处理后的维度, D 为原始轨迹维度, $\boldsymbol{R}^d$ 表示降维后的低维子空间, v_t 表示降维后的一个行为视频的轨迹特征, $\boldsymbol{V}^{\mathrm{RP}}$ 表示投影降维后所有行为视频轨迹特征的集合, $d << D$。

4 计算局部轨迹特征 v_t 被分配到第 i 个高斯单元的概率 $r(i) = \dfrac{w_i p_i(v_t)}{\sum_{i=1}^{K} w_i p_i(v_t)}$。

5 计算局部特征轨迹 v_t 关于 GMM 参数集 $\lambda = \{w_i, u_i, \sum_i\}, i = 1, \cdots, K$ 的梯度向量,归一化后级联各个梯度值,根据公式(7)-(9)求出特征轨迹的 Fisher 向量。

6 将高维的 Fisher 向量投影到一个低维子空间 $\boldsymbol{V}^{\mathrm{RP}'} = [v'_t \in \boldsymbol{R}^{d'}]$, $t = 1, \cdots, \sum_{I=1}^{m} T_I$ 中,其中, d' 表示 Fisher 向量经过随机投影二次降维后维度, $R^{d'}$ 表示二次降维后的低维子空间, v_t' 表示二次降维后的一个 Fisher 向量, $V^{RP'}$ 表示二次降维后所有 Fisher 向量的集合, $d' << d$。

7 训练 SVM 分类器,将 m 个训练视频降维编码后的轨迹特征分别贴上个对应特征行为的标签 $L = [l_1, l_2, \cdots, l_m]$, 训练出能区分不同行为动作的一个超平面。

预测部分:

8 选取 n 个测试集 $Z = [Z_1, Z_2, \cdots, Z_n, 1 \leqslant J \leqslant n]$ 中的一个新的行为视频 Z_J, 提取测试视频的轨迹特征 $\boldsymbol{Y}_J = [y_1, y_2, \cdots, y_{T_J}]^{\mathrm{T}}$, 对 $\boldsymbol{Y}_J$ 同样利用随机投影定理对其进行特征降维,将其投影到一个的低维子空间 $\boldsymbol{H}^{RP}$ 中。

9 根据步骤 5 中得到的关于 GMM 参数集,计算测试集行为视频的轨迹特征的相关的梯度向量。

10 利用随机投影定理对测试集行为视频轨迹特征的 Fisher 向量进行二次特征降维。

11 根据步骤 7 训练好的分类器对经过二次特征降维后的测试集行为视频轨迹特征的 Fisher 向量进行行为分类预测,完成行为测试集视频的识别,输出行为预测标签值。

结束

3　试　验

3.1　数据集

本节简单介绍一下试验中用到的两种数据集①,即 KTH 数据集和 UCF50 数据集②。KTH 数据集的各个动作统一在一个空旷的背景下,而 UCF50 视频则是在一个高分辨率复杂背景下完成的。

KTH 数据集包括 6 种行为动作:散步、慢跑、快跑、拳击、挥手和拍手,每一个动作是在户内、户外、户外尺寸改变、在户外搭配不同的服装 4 个不同的场景中实现的。在大部分场景中背景单一且静态,但背景噪声大。总的数据有 600 个视频样本,分辨率为 160 × 120,我们将视频样本分为 30 组,其中 5 组作为测试集,25 组为训练集分别进行试验。

UCF50 数据集有 50 个动作类,包括打篮球、跳水、高尔夫摆臂、举重、单杠、骑马等体育项目以及从 YouTube 选取的现实生活的视频片段。该数据集背景复杂,场景不一,视觉角度各异,行为识别难度较大。总的数据有 6618 个视频样本,分辨率为 320 × 240,同样我们将视频样本分为 5 组测试集,25 组训练集分别进行试验。最后我们训练和评估一个多级分类,并在所有类中将报告平均识别率作为性能衡量的标准。

① http://www.nada.kth.selcvap/actions/

② http://server.cs.ucf.edu/~vision/data/UCF50.rar

3.2 试验与试验结果分析

3.2.1 行为识别算法的准确度评估

本研究利用上述两种数据集来评估本文提出的行为识别算法。试验通过稠密采样的跟踪算法[15]和 SVM 分类器来实现人的行为识别。试验所用计算机的配置如下:内存 4GB,CPU 是 Intel Core i3 3.4GHz的台式计算机。所用代码是在 Visual Studio 2013 用 OpenCV 库开发的。两种数据集设定相同的缺省参数,对稠密轨迹跟踪算法,我们采用作者所选用的默认参数,即取 $N = 32$, 轨迹跟踪长度帧 $L = 15$, 取样步长 $W = 5$ 像素[11],随机投影中降维后的特征轨迹维度 $d = 100, d' = 48$。分类器采用 OpenCV 自带函数 SVM 来实现,其核函数类型选择 LINEAR 线性类型,算法终止条件中最大迭代次数设为 1000,精确度为 1.192092896e-07F。

本研究将数据集分为训练集和测试集,分别进行对不同算法的准确度评估。训练集是有标签学习的训练视频,测试集是用于行为分类的检测视频,二者分别从以下三个方面来进行评估:(1)不同行为动作本身的识别率;(2)BoW 模型和 GMM-FV 混合模型的性能;(3)RP 和 PCA 方法分别对行为分类的影响。试验结果如表 2 和表 3 所示。

表 2 KTH 数据集试验得出的行为识别准确率(%)

训练集行为 / 算法	快跑	挥手	拳击	慢跑	拍手	散步	平均
PCA + GMM-FV	89.3	92	89	85	89	90	89.1
PCA + BoW	81	85	90	86	84.4	88	85.7
RP + GMM-FV	89	91	92	91	90	92	90.8
测试集行为 / 算法	**快跑**	**挥手**	**拳击**	**慢跑**	**拍手**	**散步**	**平均**
PCA + GMM-FV	65	70	56	72	40	53.6	59.4
PCA + BoW	56.9	63	58.6	64	34	57	55.6
RP + GMM-FV	63	72	74	68	42	46	60.8

表 3 UCF50 数据集试验得出的行为识别准确率(%)

训练集行为 / 算法	打篮球	举重	高尔夫球摆	跳水	骑马	单杠	平均
PCA + GMM-FV	89.7	90	86	88.2	85.4	90.3	88.3
PCA + BoW	88	84	87	81.4	80	84	84.1
RP + GMM-FV	91	92	88.3	89	84.3	91	89.3
测试集行为 / 算法	**打篮球**	**举重**	**高尔夫球摆**	**跳水**	**骑马**	**单杠**	**平均**
PCA + GMM-FV	34	56	34	47.5	65	59.6	49.4
PCA + BoW	42	39	60	44	46	60.2	48.5
RP + GMM-FV	27	55	56	52	62.6	58	51.8

试验结果表明,GMM-FV 混合模型结合了空间聚合的思想,在性能和识别准确率上相比于 BoW 模型都有明显的优势。在此模型基础上利用随机投影的方法实现高维特征轨迹的降维,在两类数据集中都表现出较好识别效果。虽然 PCA 主成分分析法按照信息量贡献最大的方向进行特征降维,但是为了保证行为识别率,我们需要训练大量的行为视频,而 PCA 在处理过程中需要对一类行为动作的所有训练视频的高维特征轨迹对齐,统一实现特征降维,这样才能保证同类行为动作中主成分内在联系不会被破坏,但是大量的视频信息堆叠在一起在一定程度上增加了轨迹信息的冗余性和复杂度。而 RP 是独立地将对每一个行为视频的特征轨迹分别随机投影到一个低维子空间中,JL 引理保证了随机投影后

低维数据间的距离不变,即保证了人的行为识别的可靠性。此外由于 RP 是各个视频彼此独立进行,这就克服了 PCA 对新的测试视频进行降维时内在联系无法保证的问题。本文提出的 RP + GMM-FV 方法在 KTH 训练集和测试集试验上都表现出较好的识别率(90.8% 和 60.8%),同样,在 UCF50 数据集试验上也表现出较好的性能(训练集为 89.31%,测试集为 51.8%)。

试验中发现 KTH 和 UCF 数据集上也都出现识别率不理想的情况,如 KTH 数据集中的慢跑动作。从试验数据可以看出,在训练集上,RP + GMM-FV 试验结果中慢跑动作的识别率(91%)明显高于 PCA + GMM-FV 中的试验效果(85%),但在测试集部分 RP 的识别率(68%)却不及 PCA(72%)测试效果好。分析表明,产生这一现象的原因是由于 RP 随机投影的原理是在保证原始的输入信号间距离不变的基础上将这些输入信号投影到一个低维子空间中,而为了避免不同行为动作在视频中的位置或角度的不同对试验结果的影响,我们对轨迹提取的位置信息进行了归一化处理。这样慢跑相对快跑动作除了在时间域上有明显区别外,在空间结构上差异性就不够突出,因而试验表明 RP 随机投影在相似的空间行为动作上(如快跑和慢跑)和 PCA 相比并不能表现出明显的优势。再如 UCF50 数据集中的打篮球动作,对其识别的准确率低于 30%。产生这一现象的原因主要是在 UCF50 数据集中,各个视频动作的背景比较复杂,部分视频摄像机窗口尺寸会大幅度地缩小或放大,这样摄像机在跟随目标移动的过程中会造成干扰和背景大幅度改变,影响了行为轨迹的正常提取,这在很大程度上就降低了最终的行为识别的准确度。因而该行为识别算法在动态背景迅速移动的情况下存在一定的缺陷。

3.2.2　行为识别算法的效率评估

本节利用上述两种数据集来评估本文提出的算法在行为识别上的效率。试验运行在 Matlab R2012b 上,在 KTH 和 UCF50 两种数据集上分别来比较 RP 随机投影定理和 PCA 主成分分析方法的运行时间,以评估二者的效率。试验分别从两种数据集中各选择三种行为识别动作(快跑,拳击,挥手,打篮球,举重和高尔夫球摆),每种行为动作各选择 30 个视频的轨迹信息进行特征降维。从前文可知,这些原始的特征轨迹信息的维度均为 2010 维。试验中,我们分别用 RP 和 PCA 两种方法将这些高维特征轨迹的维度分别降到 10,30,60,90,120,150 维来比较二者的计算时间(单位:s)。其中 RP 是对同一类行为动作的每一个视频提取出的特征轨迹分别进行投影降维,而 PCA 是对同一个行为动作的所有视频的特征轨迹对齐后,进行统一特征降维。试验记录两种数据集上各自程序运行的时间,结果如表 4,图 2 和图 3 所示。

表 4　RP 和 PCA 两种方法的降维时间比较(单位:s)

		KTH			UCF50		
		挥手	快跑	拳击	打篮球	举重	高尔夫球摆
10 维	RP	0.242	0.257	0.304	0.429	0.516	0.332
	PCA	21.87	31.29	36.45	468.72	2312.94	47.12
30 维	RP	0.243	0.291	0.334	0.469	0.554	0.343
	PCA	24.65	30.00	42.80	686.05	2629.56	56.91
60 维	RP	0.251	0.310	0.415	0.561	0.568	0.398
	PCA	32.96	41.38	52.22	678.64	3021.24	70.69
90 维	RP	0.277	0.339	0.469	0.582	0.668	0.408
	PCA	39.06	51.19	65.11	767.94	3167.43	87.55
120 维	RP	0.305	0.360	0.508	0.664	0.695	0.454
	PCA	46.26	57.95	75.86	796.76	3324.37	97.46
150 维	RP	0.353	0.410	0.586	0.623	0.846	0.698
	PCA	52.15	65.63	89.18	824.63	3510.58	118.35

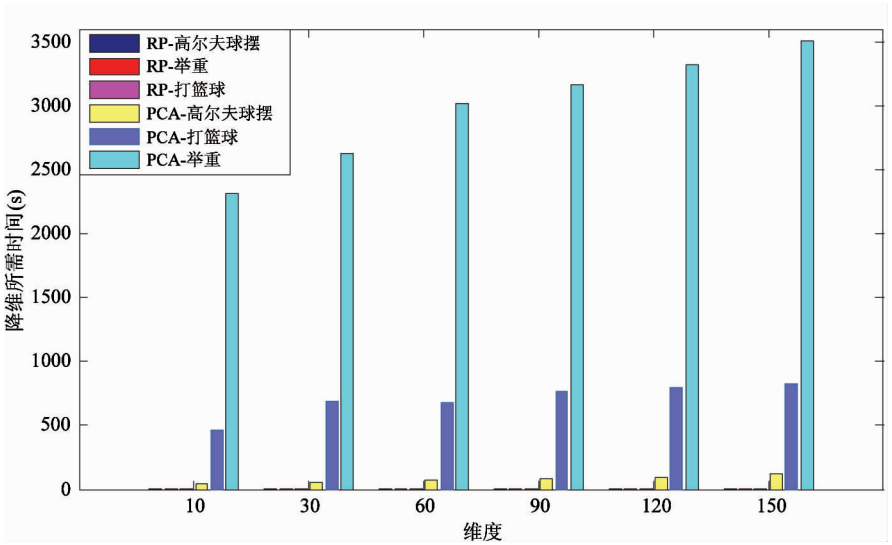

图 2　UCF50 数据集试验中 RP 和 PCA 两种方法的降维时间比较

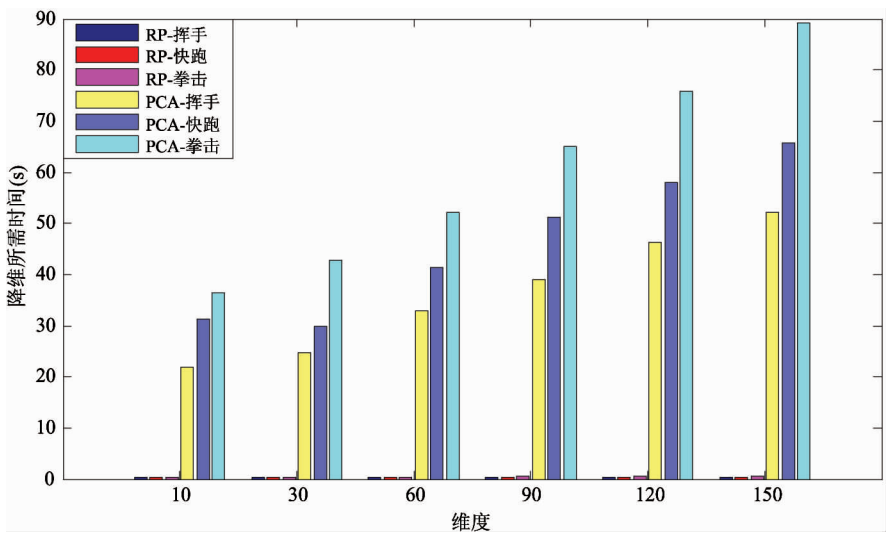

图 3　KTH 数据集试验中 RP 和 PCA 两种方法的降维时间比较

从图 2 、图 3 中可以看出，RP 在 KTH 和 UCF50 数据集试验中都明显提高了行为识别算法的效率。在 KTH 数据集中，行为动作简单，背景单一，提取轨迹条数比较均匀（每个动作平均轨迹条数为 5000），RP 降维处理节约了计算时间。在 UCF50 数据集中，由于人的动作和背景比较复杂，故提取的人的行为特征轨迹数量大（如举重动作，轨迹条数高达 15000 以上），在此数据集上采用 PCA 降维运行时间和 KTH 数据集相比耗时很长，而经过 RP 降维处理的时间在两个数据集中的表现并没有很大差别。试验结果表明，相对于 RP，PCA 降维需要进行矩阵的特征分解，计算代价高，会随着数据量的增大而影响运行效率；而 RP 降维则仅需要进行简单的矩阵运算，特别是可以构造出稀疏的满足 RIP 性质的随机矩阵，进而大大提高了算法的运行效率。

3.2.3　行为识别算法的综合评估

本文所提出的人的行为识别方法在准确度和效率上均有较好的改进。从表 2 和表 3 可以看出，GMM-FV 混合模型结合了空间聚合的思想，在相同降维方法（PCA）条件下，其行为识别的精度相比 BoW 模型有很大的提高；此外，分析试验结果发现，在相同的 GMM-FV 混合模型下，RP 和 PCA 两种降维方法对本文提出的人的行为识别率的影响几乎不大（近似或 PCA 略高于 RP）。从表 4 中可以明显看

出,在保证行为识别精度基本一致的基础上,RP 在面对两种数据集降维时所用的时间差异不大,这样对处理大量高维数据是非常有利的。RP 大大节约了该行为算法运行的时间,提高了算法的效率。为了整体体现本文算法的优越性,以下给出该算法在 KTH 和 UCF50 两种数据集上的性能比较(见表 5)。其中降维所需的时间是指本文算法中特征轨迹降到 $d = 100$ 维时所需的时间。

表 5　行为识别算法在 KTH 和 UCF50 两种数据集上的性能比较

	行为识别准确度(%)	降维所需时间(s)	
		PCA	RP
KTH	75.8	63.16	0.39
UCF50	70.6	1506.23	0.67

4　结　论

本文提出的是一种基于随机投影和 Fisher 向量的人的行为识别方法,试验结果表明该方法能够有效提高运行效率,降低训练器计算复杂度,提高行为识别的准确率。在以后的工作中将考虑利用前景背景分离的方法来实现视频动态背景的去除,以解决前面遇到的相机移动问题[23]。此外本文提出的基于人的特征行为的低维表示方法,既可以利用上述 SVM 分类器进行学习和训练,也可以从深度学习的角度出发,如卷积神经网络(CNNs)或深置信网络(DBNs)等学习方法,以达到我们所期待的更理想的实验效果。

参考文献

[1] 徐勤军,吴镇扬. 视频序列中的行为识别研究进展. 电子测量与仪器学报, 2014,28(4): 343-351

[2] Poppe R. A survey on vision-based human action recognition. *Image and vision computing*, 2010, 28(6): 976-990

[3] 申利民, 刘称称, 尤殿龙. 一种面向自动化设备的行为监测与异常诊断方法. 小型微型计算机系统, 2015, 36(1): 126-132

[4] Masoud O, Papanikolopolulos N. A method for human action recognition. *Image and Vision Computing*, 2003, 21(8): 729-743

[5] Scovanner P, Ali S, Shan M. A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 2007. 357-360

[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005. 1, 886-893

[7] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies. In: Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Alaska, USA, 2008. 1-8

[8] Nowozin S, Bakir G, Tsuda K. Discriminative subsequence mining for action classification. In: Proceedings of the IEEE 11th International Conference on Computer Vision, Rio De Janeiro, Brazil, 2007. 1-8

[9] Tirilly P, Claveau V, Gros P. Language modeling for bag-of-visual words image categorization. In: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, Melbourne, Australia, 2008. 249-258

[10] Guha T, Wadr K. Learning sparse representations for human action recognition. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on*, 2012, 34(8): 1576-1588

[11] Bomma S, Favaro P, Robertson N M. Sparse representation based action and gesture recognition. In: Proceedings of the 20th IEEE International Conference on Image Processing (ICIP), Melbourne, Australia, 2013. 141-145

[12] Bingham E, Mannila H. Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining, San Francisco, USA, 2001. 245-250

[13] Fern X Z, Brodley C E. Random projection for high dimensional data clustering: A cluster ensemble approach. In: Proceedings of the 20th International Conference on Machine Learning, Washington, USA, 2003. 3: 186-193

[14] Perronnin F, Sanchez J, Mensink T. Improving the fisher kernel for large-scale image classification. In: Computer

Vision - ECCV 2010. Berlin/Heidelberg: Springer, 2010. 143-156

[15] Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 2013. 3551-3558

[16] Wang H, Klaser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 2013, 103(1): 60-79

[17] 方红, 章权兵, 韦穗. 基于非常稀疏随机投影的图像重建方法. 计算机工程与应用, 2007, 43(22): 25-27

[18] Achlioptas D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, 2003, 66(4): 671-687

[19] Zhang K, Zhang L, Yang M H. Real-time compressive tracking. In: Computer Vision - ECCV 2012. Berlin/Heidelberg: Springer, 2012. 864-877

[20] Candes J, Tao T. Decoding by linear programming. *Information Theory*, *IEEE Transactions on*, 2005, 51(12): 4203-4215

[21] Sanchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 2013, 105(3): 222-245

[22] Huang Y, Englehart K B, Hudgins B, et al. A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. *Biomedical Engineering*, *IEEE Transactions on*, 2005, 52(11): 1801-1811

[23] Berger M, Seversky L M. Subspace tracking under dynamic dimensionality for online background subtraction. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA, 2014. 1274-1281

A human action recognition method based on random projection and Fisher vector

He Jun, Xue Ying, Hu Zhaohua, Sun Wei

(Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science and Technology, Nanjing 210044)

Abstract

To improve the effectiveness of video-based human action recognition, the study proposed a new action recognition method using the random projection (RP) and Fisher vector (FV) on the basis of the definition and analysis of the Gaussion mixture model (GMM). The method projects high dimensional trajectory descriptors into a low dimensional subspace through random projection to realize the dimension reduction for feature trajectory, then uses the GMM-FV model to perform the spatial clustering coding for the trajectory feature vector after dimension reduction to improve the recognition accuracy, and finally, uses the random projection again to secondly reduce the dimension of the Fisher coding vector to reduce the computation complexity. The experiment performed using the datasets of KTH and UCF50 showed that compared with the existing recognition algorithms, the proposed method had the lower computation complexity, the higher recognition accuracy, and the good recognition robustness in the experiment using the two datasets.

Key words: action recognition, feature trajectory, random projection (RP), feature dimension reduction, Fisher vector (FV)