局部分布信息增强的视觉单词描述与动作识别

张　良*　　鲁梦梦　　姜　华
(中国民航大学智能信号与图像处理天津市重点实验室　天津　300300)

摘　要：传统的单词包(Bag-Of-Words, BOW)算法由于缺少特征之间的分布信息容易造成动作混淆，并且单词包大小的选择对识别结果具有较大影响。为了体现兴趣点的分布信息，该文在时空邻域内计算兴趣点之间的位置关系作为其局部时空分布一致性特征，并提出了融合兴趣点表观特征的增强单词包算法，采用多类分类支持向量机(Support Vector Machine, SVM)实现分类识别。分别针对单人和多人动作识别，在 KTH 数据集和 UT-interaction 数据集上进行实验。与传统单词包算法相比，增强单词包算法不仅提高了识别效率，而且削弱了单词包大小变化对识别率的影响，实验结果验证了算法的有效性。

关键词：人体行为识别；局部分布特征；增强单词包模型；支持向量机
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2016)03-0549-08
DOI: 10.11999/JEIT150410

An Improved Scheme of Visual Words Description and Action Recognition Using Local Enhanced Distribution Information

ZHANG Liang　　LU Mengmeng　　JIANG Hua
(*Key Laboratory of Advanced Signal and Image Processing*, *Civil Aviation University of China*, *Tianjin* 300300, *China*)

Abstract: The traditional Bag-Of-Words (BOW) model easy causes confusion of different action classes due to the lack of distribution information among features. And the size of BOW has a large effect on recognition rate. In order to reflect the distribution information of interesting points, the position relationship of interesting points in local spatio-temporal region is calculated as the consistency of distribution features. And the appearance features are fused to build the enhanced BOW model. SVM is adopted for multi-classes recognition. The experiment is carried out on KTH dataset for single person action recognition and UT-interaction dataset for multi-person abnormal action recognition. Compared with traditional BOW model, the enhanced BOW algorithm not only has a great improvement in recognition rate, but also reduces the influence of BOW model's size on recognition rate. The experiment results of the proposed algorithm show the validity and good performance.

Key words: Human action recognition; Local distribution features; Enhanced Bag-Of-Words (BOW) model; Support Vector Machine (SVM)

1　引言

人体行为识别是计算机视觉领域的热门研究课题之一，它具有非常重要的现实意义，在智能视频监控、虚拟现实、医疗辅助和运动员动作分析等方面[1]有着十分广泛的应用。但是，由于背景复杂、摄相机抖动、光照变化、遮挡以及不同动作者的类内差异等都使目前的行为识别面临着很大的挑战。

基于局部时空特征的单词包(Bag-Of-Words, BOW)算法作为一种简单有效、鲁棒性较强的行为表示方法，在行为识别领域得到了广泛的应用。传统单词包算法[2−6]是对检测到的时空兴趣点计算表观特征描述子，并对描述子聚类形成视觉单词包。通过将每个视频的特征向量映射到单词包空间得到视频的词频分布直方图表示。文献[7]采用 BOW 算法和状态空间法进行行为识别，在计算量和精确度之间实现了平衡。文献[8]从全局的角度出发，筛选时空单词构建显著性视觉词汇，弥补了基于局部区域的单词包算法造成的全局信息的缺失。文献[9]选取单词包中包含信息量丰富的视觉单词构建新的视觉单词包，使最终的视觉词汇更具有效性。

但是传统的单词包算法只考虑了每个视觉词汇出现的次数，而忽略了特征之间的位置关系。近年来，许多研究者开始把特征之间的时空关系作为人体行为识别的关键。文献[10]提出了情境感知的时空描述子，每种行为由特定的动作单元表示，并且涵

收稿日期：2015-04-08；改回日期：2015-12-08；网络出版：2016-01-22
*通信作者：张良　stonemark@vip.163.com
基金项目：国家自然科学基金(61179045)
Foundation Item: The National Natural Science Foundation of China (61179045)

盖了动作单元之间的几何信息，弥补了传统 BOW 造成的语义鸿沟问题。文献[11]计算特征点出现的先后次序，将代表词汇出现先后次序的二值矩阵用于行为识别。 文献[12]基于单词包模型，采用高阶共生矩阵将视觉单词映射到共生空间对行为进行表示，识别结果受单词包大小的影响减弱。文献[13]通过分析视觉单词之间的共生关系，选择稳定的兴趣点用于行为识别。文献[14]采用直方图交叉核计算词汇之间的时空上下文关系，对视觉词汇进行加权实现行为识别。

考虑到同种动作拥有相似的兴趣点分布信息，不同动作的兴趣点的分布情况不相同，本文提出了融合兴趣点的表观特征和局部时空分布一致性特征的增强单词包算法进行行为识别。该算法不需要精确的人体检测定位、背景减除、目标跟踪等，也不要求特征点在全局空间严格匹配，计算过程简便。主要包括以下 3 个方面：(1)采用 Harris3D 检测器提取视频中的时空兴趣点，确保特征点的有效性；(2)提出了构建兴趣点的位置关系作为局部时空分布一致性特征，它是直接对检测到的特征点进行计算而非时空单词，避免了映射到单词包空间带来的量化误差。然后结合兴趣点的 HOG3D 特征构建增强单词包模型，此模型不仅包含了表观信息，也涵盖了空间分布信息；(3)行为建模与分类，采用多类分类 SVM 实现分类识别。通过实验证明，增加局部时空分布特征的单词包模型能够弥补传统单词包算法因缺少词汇之间位置关系而造成的缺陷，提高了识别精度，减少动作之间的混淆程度。

2 特征检测与描述

本文提出了增强单词包算法的人体行为识别，图 1 给出了本文的算法框架，大体分为两个部分：训练部分和测试部分。在训练阶段，首先采用 Harris3D 检测器提取训练视频中的局部时空兴趣点，然后计算兴趣点的表观特征集和局部时空分布特征集。分别对这两类特征集聚类得到表观特征单词包和局部分布特征单词包。基于这两类单词包，统计每个单词在训练视频中出现的次数，得到训练视频的表观特征直方图和局部分布特征直方图，将这两个直方图向量级联起来并分配动作类别编号，作为训练视频的类别统计直方图。它是一个 K 维的特征向量，K 值由聚类中心个数确定。计算所有训练视频的类别统计直方图，输入 SVM 训练，确定动作分类器模型。在测试部分，以同样的方法计算测试视频中的时空兴趣点以及两类特征集，采用 KNN 算法，将特征集投影到单词包空间，统计每个视觉单词出现的次数得到测试视频的词频分布直方图，输入训练好的 SVM 进行识别。

2.1 Harris3D 时空兴趣点检测

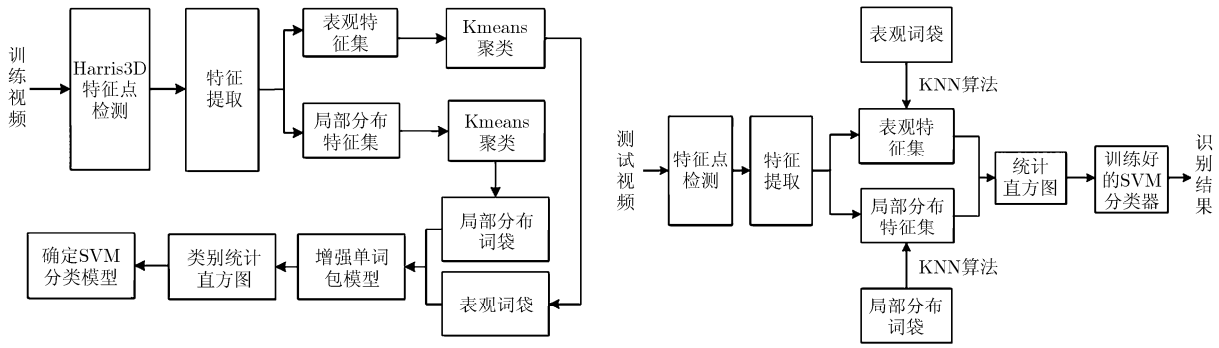
时空兴趣点是典型的局部特征，反映了视频中在时间域和空间域都有明显变化的地方，具有很好的特征描述和类别区分能力。目前应用比较广泛的是文献[15]提出的 Harris3D 时空兴趣点。

首先，将视频序列 f 变换到 3 维的高斯空间，如式(1)所示。

$$L\left(\cdot;\sigma_l^2,\tau_l^2\right)=g\left(\cdot;\sigma_l^2,\tau_l^2\right)*f(\cdot) \tag{1}$$

其中， $g\left(x,y,t;\sigma_l^2,\tau_l^2\right)=\dfrac{1}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}}\mathrm{e}^{\left[-\left(x^2+y^2\right)/2\sigma_l^2-t^2/2\tau_l^2\right]}$ 为 3D 高斯核函数， σ_l^2， τ_l^2 分别为互相独立的空间尺度因子和时间尺度因子。在给定尺度因子 σ_l^2， τ_l^2 后 ，采用由 L 的一阶导数组合而成的二阶矩矩阵和一个 3D 高斯窗口的卷积来寻找兴趣点。如式(2)所示。

$$\boldsymbol{\mu}=g\left(\cdot;\sigma_i^2,\tau_i^2\right)*\begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \tag{2}$$



(a)行为识别训练部分 (b)行为识别测试部分

图 1 行为识别框架

其中，$\sigma_i^2 = s\cdot\sigma_l^2$，$\tau_i^2 = s\cdot\tau_l^2$ 分别为空间尺度因子 σ_l^2 和时间尺度因子 τ_l^2 的积分尺度，$L_x\left(\cdot;\sigma_l^2,\tau_l^2\right) = \dfrac{\partial(g*f)}{\partial x}$ 为在 x 方向上的一阶导数，L_y，L_t 计算方法与之相似。因此，$\boldsymbol{\mu}$ 的特征值 λ_1，λ_2，λ_3 对应序列 f 的 3 个方向 (x,y,t) 的变化。角点响应函数为

$$\begin{aligned} H &= \det(\boldsymbol{\mu}) - k\times\mathrm{trace}^3(\boldsymbol{\mu}) \\ &= \lambda_1\lambda_2\lambda_3 - k\left(\lambda_1+\lambda_2+\lambda_3\right)^3 \end{aligned} \tag{3}$$

把角点响应函数取正最大值的像素点作为时空兴趣点。由于每个动作都有一定的持续时间，为了估计行为的时空范围，文献[15]引入了时空高斯块，其定义为

$$f\left(x,y,t;\sigma_0^2,\tau_0^2\right) = \frac{1}{\sqrt{(2\pi)^3\sigma_0^4\tau_0^2}}\mathrm{e}^{\left[-\left(x^2+y^2\right)/2\sigma_0^2 - t^2/2\tau_0^2\right]} \tag{4}$$

将表示此时空事件的高斯块映射到高斯空间，得到其线性空间为

$$L\left(x,y,t;\sigma^2,\tau^2\right) = g\left(x,y,t;\sigma^2,\tau^2\right) * f\left(x,y,t;\sigma_0^2,\tau_0^2\right) \tag{5}$$

根据高斯函数的性质，可以得到函数 L 的规范形式：

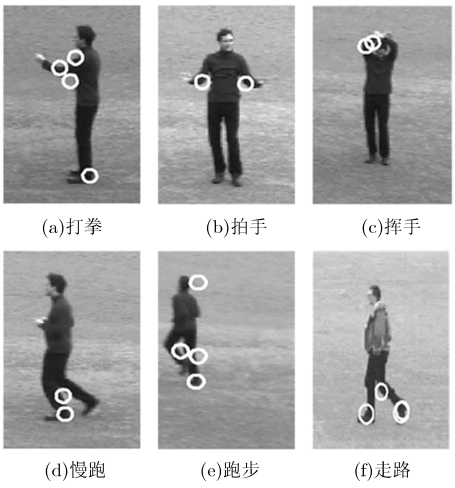
$$L\left(x,y,t;\sigma^2,\tau^2\right) = g\left(x,y,t;\sigma^2+\sigma_0^2,\tau^2+\tau_0^2\right) \tag{6}$$

分别求 L 在 x, y, t 方向的二阶导数 L_{xx}，L_{yy}，L_{tt}，规范化得到 $L_{xx,\mathrm{norm}}$，$L_{yy,\mathrm{norm}}$，$L_{tt,\mathrm{norm}}$，规范化后函数极值在高斯块的中心取得，即 $\sigma^2 = \sigma_0^2$，$\tau^2 = \tau_0^2$。然后计算归一化的时空拉普拉斯算子 $\nabla^2_{\mathrm{norm}}L$ 同时在时间尺度和空间尺度上的极值点，得到的即为最佳时间尺度和空间尺度。$\nabla^2_{\mathrm{norm}}L$ 的计算公式为

$$\begin{aligned} \nabla^2_{\mathrm{norm}}L &= L_{xx,\mathrm{norm}} + L_{yy,\mathrm{norm}} + L_{tt,\mathrm{norm}} \\ &= \sigma^2\tau^{1/2}\left(L_{xx}+L_{yy}\right) + \sigma^2\tau^{3/2}L_{tt} \end{aligned} \tag{7}$$

为了使兴趣点的尺度刚好匹配相应行为的时空范围，实验中进行了尺度自适应迭代。

Harris3D 检测器在 KTH 数据库上检测到的兴趣点如图 2 所示：



(a)打拳　(b)拍手　(c)挥手

(d)慢跑　(e)跑步　(f)走路

图2 特征点在KTH视频帧上的展示

2.2 特征描述子

本文提出了融合兴趣点的表观特征和局部时空分布特征的行为描述子。其中表观特征采用 3 维梯度方向直方图[16](Three Dimension Histogram of Gradients, HOG3D)，它代表了兴趣点邻域内的像素变化信息；局部时空分布特征代表了时空邻域内兴趣点之间的位置分布信息。

2.2.1 HOG3D 描述子　HOG3D 描述了特征点邻域的平均梯度，它的计算主要包括 3 个方面：梯度的计算、梯度方向的量化、直方图的计算以及描述子的生成。

(1)梯度计算：给定一个视频序列 $f(x,y,t)$，其一阶导数分别为 f_x，f_y，f_t。其中，f_x 所对应的积分视频描述为：$if_x(x,y,t) = \sum_{x'\le x,y'\le y,t'\le t} f_x(x',y',t')$，$if_y$，$if_t$ 的定义与 if_x 类似。在视频中任一兴趣点 $(x,y,t)^{\mathrm{T}}$ 处，它所在的时空邻域 $\boldsymbol{b} = (x,y,t,w,h,l)^{\mathrm{T}}$，其中 w, h, l 分别为立方体的宽度、高度和长度。其平均梯度表示为：$\overline{\boldsymbol{g}}_{\boldsymbol{b}} = \left(\overline{g}_{bx},\overline{g}_{by},\overline{g}_{bt}\right)^{\mathrm{T}}$，其中

$$\begin{aligned} \overline{g}_{bx} = &\left[if_x(x+w,y+h,t+l) - if_x(x,y+h,t+l)\right. \\ &\left.-if_x(x+w,y,t+l) + if_x(x,y,t+l)\right] \\ &-\left[if_x(x+w,y+h,t) - if_x(x,y+h,t)\right. \\ &\left.-if_x(x+w,y,t) + if_x(x,y,t)\right] \end{aligned} \tag{8}$$

(2)梯度方向量化：采用正二十面体进行量化。设 $\boldsymbol{p} = (p_1,p_2,\cdots,p_{20})$ 为正二十面体各个面的中心，对每个时空兴趣点的平均梯度 $\overline{\boldsymbol{g}}_b$ 在各个坐标轴上进行投影，得到投影向量 $\hat{\boldsymbol{q}}_{\boldsymbol{b}}$：

$$\hat{\boldsymbol{q}}_{\boldsymbol{b}} = \left(\hat{q}_{b1},\hat{q}_{b2},\cdots,\hat{q}_{bm}\right)^{\mathrm{T}} = \frac{P\cdot\overline{g}_b}{\left\|\overline{g}_b\right\|_2} \tag{9}$$

其中，$\hat{\boldsymbol{q}}_{\boldsymbol{bi}} = \left\|p_i\right\|_2\cdot\cos_{\angle}\left(p_i,\overline{g}_b\right) = \left\|\overline{g}_b\right\|_2^{-1}\cdot p_i^{\mathrm{T}}\cdot\overline{\boldsymbol{g}}_{\boldsymbol{b}}$。对投影向量 $\hat{\boldsymbol{q}}_{\boldsymbol{b}}$ 阈值化得到 $\hat{\boldsymbol{q}}_{\boldsymbol{b}}'$，投影梯度为 $\boldsymbol{q}_{\boldsymbol{b}} = \dfrac{\left\|\overline{g}_b\right\|_2\cdot\hat{\boldsymbol{q}}_{\boldsymbol{b}}'}{\left\|\hat{\boldsymbol{q}}_{\boldsymbol{b}}'\right\|_2}$。

(3)直方图计算及描述子生成：在时空兴趣点 $\boldsymbol{p} = (x_p,y_p,t_p,w_p,h_p,l_p)$ 处，$w_p\times h_p\times t_p$ 是兴趣点周围所选视频块的大小。将所选的视频块进一步细分为 $s\times s\times s$ 个小的视频块 b_i，每个小视频块的平均梯度为 $\mathrm{g}\,\overline{g}_{bi}$，经过投影之后得到投影向量 $\boldsymbol{q}_{\boldsymbol{bi}}$。对各个视频块 b_i 经过量化之后的平均梯度向量 $\boldsymbol{q}_{\boldsymbol{bi}}$ 进行求和得到点 p 的特征向量直方图 h_p。在任意尺度的时空特征点 (x_s,y_s,t_s) 处，其所在视频块的特征向量都可以通过这样的划分方式求得固定维数的特征向量。计算每个小区域的梯度向量，并量化合并成一个单一的特征向量 $\boldsymbol{d}_{\boldsymbol{s}} = (d_1,d_2,\cdots,d_{M^2N})$ 作为点 (x_s,y_s,t_s) 处的表观特征描述子。

2.2.2 局部时空分布一致性特征　局部时空分布一致性特征描述了兴趣点的分布信息。以 KTH 数据

库为例，任意选取 KTH 中的两个人(p1, p2)，分别执行 4 种不同的动作(跑步，慢跑，打拳，拍手)，计算相应的时空兴趣点。由图 3 可以看出，对于同种动作而言，尽管产生的时空兴趣点不尽相同，但是兴趣点的分布在其局部空间内保持一致，把这一特性作为兴趣点的局部时空分布一致性特征并用于行为分类识别。

对于任意一兴趣点 $p(x,y,t)$，其局部时空分布特征计算如下：

(1)选取兴趣点 p 的时空邻域块大小，统计邻域块内兴趣点的个数，并记为 n。

(2)记录兴趣点之间的位置关系，包含距离和方向，以向量形式表示。对于邻域块中的任一兴趣点 $s(x_i,y_i,t_i)$，其与兴趣点 p 之间的位置关系记作 $\boldsymbol{r_i}=(x-x_i,y-y_i,t-t_i)$，$i=1,2,\cdots,n-1$。

(3)计算兴趣点 p 与时空邻域内其余 $(n-1)$ 个兴趣点之间的相对位置关系，统计即可获得兴趣点 p 的局部时空分布特征描述子 $\boldsymbol{r}=(r_1,r_2,\cdots,r_{n-1})$，且 $\|r_1\|_2>\|r_2\|_2\cdots>\|r_{n-1}\|_2$。

考虑到距离 p 越近的兴趣点对兴趣点 p 的描述贡献越大，为了简化计算，选取距离 p 最近的前 $n/3$ 个兴趣点，并计算这 $n/3$ 个兴趣点与 p 的时空位置关系，即把 $(r_1,r_2,\cdots,r_{n/3})$ 作为兴趣点 p 的局部时空分布特征。局部时空相对位置信息的计算是基于特征点本身，不会引起由特征点映射到单词包空间[11,12]造成的量化误差，也不需要其它预处理操作，计算方便的同时也获得了充足的分布信息，增强了描述子的鲁棒性。

3 行为建模和分类

虽然同类动作产生的时空兴趣点不尽相同，但是同类动作具有相似的表观特征集和时空分布特征集。通过增强单词包模型构建动作原型，以原型来描述视频中的行为，采用 SVM 分类器实现分类识别。

增强单词包模型是基于 K 均值聚类，通过分别对归一化后的局部时空分布特征集和表观特征集进行聚类得到时空分布单词包 $\boldsymbol{L}=(l_1,l_2,\cdots,l_m)$ 和表观特征单词包 $\boldsymbol{F}=(q_1,q_2,\cdots,q_n)$，$m$, n 为聚类中心的大小，l_i 为时空分布单词，q_i 为表观特征单词。将两个单词包进行串联融合得到增强后的单词包，其大小为 $(m+n)$。

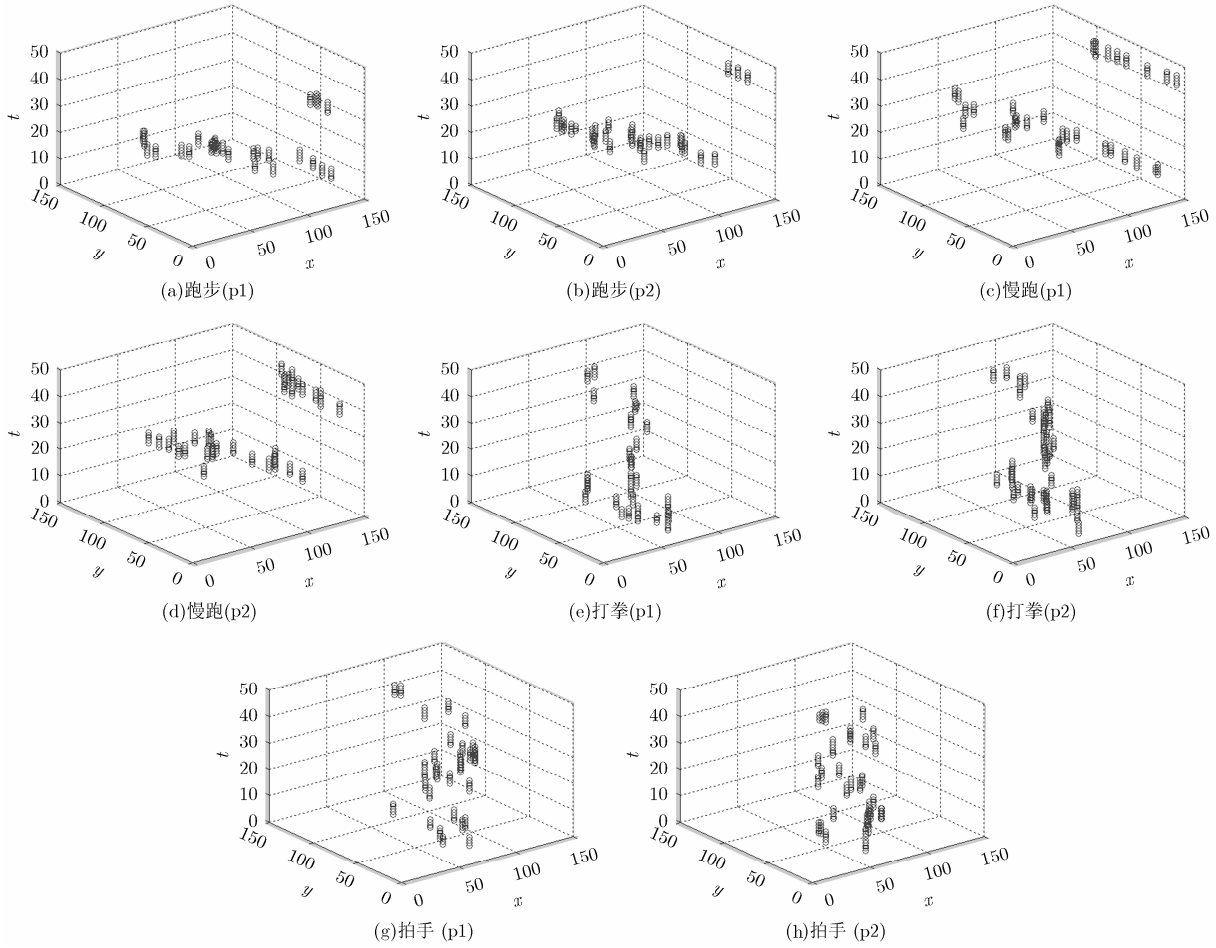


图 3 不同动作兴趣点的时空分布

对于每一个视频，首先计算视频的兴趣点集 $\boldsymbol{p}=(p_1,p_2,\cdots,p_s)$，然后分别计算每个兴趣点 p_i 的表观特征描述子和局部时空分布特征描述子。采用 KNN 算法计算这两类特征描述子与对应单词包中的每个单词之间的距离，将其归类为距离最近的单词类别。接着将视频中的行为看作是这两个单词包中的单词共同构成的一个文本，统计增强单词包中每个单词出现的次数，即可得到表示此行为的词频分布直方图 $\boldsymbol{H}=(h_1,h_2,\cdots,h_{m+n})$，其中 h_i 代表视频中第 i 个时空单词出现的频率 $(i=1,2,\cdots,m+n)$。

在分类阶段采用基于径向基核函数的多类分类 SVM。对于一个包含 n 类动作类别的数据集，设类别标签为 $\boldsymbol{C}=(c_1,c_2,\cdots,c_n)$，$c_i$ 表示动作类别 i。在训练阶段，将每个视频中的词频分布直方图 $\boldsymbol{H}$ 和类别标签 c_i 输入 SVM 进行训练，得到每个行为的原型；在测试阶段，计算测试视频的特征描述子，并映射到增强单词包空间获得其词频分布直方图 $\boldsymbol{H}'$，将 $\boldsymbol{H}'$ 输入已训练的 SVM，输出结果 c_i 即为测试视频的动作类别 i。

4　实验及结果分析

4.1 单人行为识别实验及结果分析

实验采用的单人动作数据集是 KTH 数据集。它包含了 6 种动作：走路，慢跑，跑步，打拳，挥手，拍手，由 25 个动作者在 4 种不同的场景(室外，有尺度变化的室外，动作者有着装变化的室外，室内)下完成，共计 600 个视频。

4.1.1 传统单词包算法的单人行为识别　传统的 BOW 算法对表观特征描述子进行 K 均值聚类，聚类中心个数的选择不仅代表了单词包的大小，而且也决定着识别效果的好坏。实验中依次选择单词包的大小为 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700 和 800。如图 4 所示，通过多次试验我们得到下面的结论：当 K=300 时识别效果较好。当 K 在 100～300 时，随着单词包的增大，识别效率有所提高。但是当单词包大小超过 300 时，反而降低了识别效率。

在 KTH 数据集中，采用留一交叉验证，把 24 位动作者的视频数据当作训练集，剩余 1 个人的全部行为视频当作测试集。重复此过程 10 次，统计每个动作被正确识别和错误识别的次数，得到 KTH 的混淆矩阵如图 5 所示。从混淆矩阵可以得到基于传统 BOW 算法的多类分类 SVM 行为识别方法在在 KTH 数据集上的平均识别率为 87.7%。

4.1.2 增强单词包算法的单人行为识别　为了有效描述特征点的局部分布特征，选取的局部邻域块的大小为 30×15×4，在此时空体内计算特征点之间的分布信息。考虑到单词包大小会影响识别效果，为了选取最佳的特征单词包，进行多次试验，分别依次选择表观单词包和局部时空分布单词包的大小为 100, 200, 300, 400, 500, 600, 700 和 800，实现(100, 100) ,⋯, (100, 800),⋯,(800,100),⋯,(800, 800)共计 64 种单词包组合。通过多次实验，统计在每种单词包组合下的识别效率。如图 6 所示，“*”所在位置的识别效果最好，即最佳单词包大小为(400, 400)。并且可以看出，随着单词包大小的增加，识别率有所提高，但是当单词包大小超过一定范围时，识别率基本保持稳定。说明增强单词包算法减弱了识别率对单词包大小变化的敏感程度。

同样采用留一交叉验证，重复实验 10 次，将 10 次的平均值作为最终的分类精度。所得结果的混淆矩阵如图 7 所示。计算可知，增强单词包算法在 KTH 数据集上的平均识别率为 93.6%。与传统单词包算法相比，不仅识别率有了很大提高，并且减弱了单词包大小变化对识别率的影响。

从表 1 可以看出，与文献[2,3,16]中基于传统单词包算法的行为识别相比，本文提出的增强单词包算法算法在 KTH 数据集上识别效果更好，与文献[12,13,14]中结合单词的时空上下文信息的行为识别方法相比，本文也取得了较为满意的识别效果。

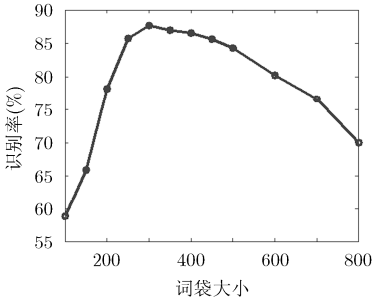


图 4 单词包大小对 KTH 识别率的影响

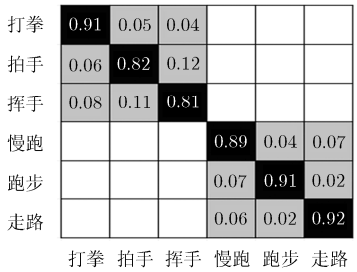


图 5 KTH 数据库的混淆矩阵

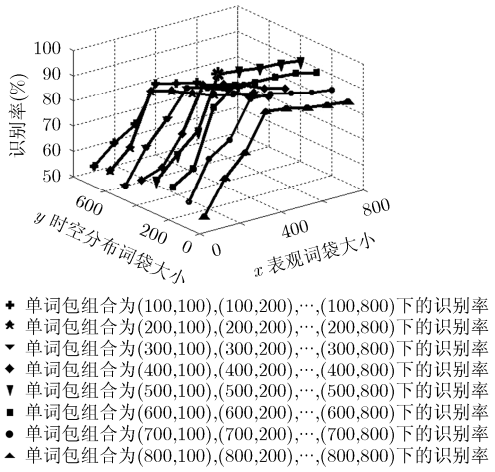


图 6 增强单词包大小对 KTH 识别率的影响

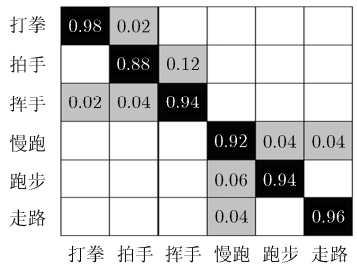


图 7 KTH 数据库的混淆矩阵

表 1 与现有方法识别率对比(%)

算法	KTH 数据库
本文算法	**93.6**
文献[2]	93.3
文献[3]	90.0
文献[12]	91.2
文献[13]	90.5
文献[14]	93.0
文献[16]	91.4

4.2 两人交互异常行为识别实验及结果分析

UT-interaction 数据库一共包含 6 种交互行为类别，包括拥抱、踢打、用手指人、推打、挥手、捶打。实验时将每个视频数据截断为 10 s 左右的视频片段，一共得到 120 个视频片段。并将这 6 种行为分为两类：异常行为(推打、踢打、捶打)和正常行为(拥抱、握手、用手指人)。

增强单词包算法在 KTH 数据库上的成功应用证明了算法的有效性，因此将其应用在两人交互异常行为识别中。两人交互行为中往往包含许多单一背景且只有一个目标出现，或者是两目标之间距离很大的视频帧，这些都属于无关帧。为了避免对无关视频帧的计算，在提取兴趣点之前采取一些预处理措施来选取感兴趣的帧。

图 8(a)是采用高斯混合模型(GMM)提取到的前景图像，图 8(b)是对前景图像进行形态学处理得到的增强后的前景图。并把前景目标的中心看作运动目标所在的位置，如图 9 所示。

感兴趣帧的选取是以前景目标之间的位置关系为基础，分为以下两种情况：

(1)交互过程中两人没有出现遮挡：计算前景目标之间的距离，将两目标之间的距离小于给定阈值的视频帧作为候选的感兴趣帧。

(a)前景图像 (b)增强后前景图像

图 8 前景检测

图 9 目标中心标记

(2)交互过程中出现遮挡：当两目标首次出现距离小于阈值时，记录时刻 $T1$，接着两目标出现遮挡。当两个目标又重新分开，且在 $T2$ 时刻两目标之间的距离小于给定阈值，在 $T2+\tau$ 时刻两目标之间的距离大于给定的阈值时，记录时刻 $T2$，并将 $T1$-$T2$ 时间段内的视频帧记为感兴趣帧，τ 为帧间隔。

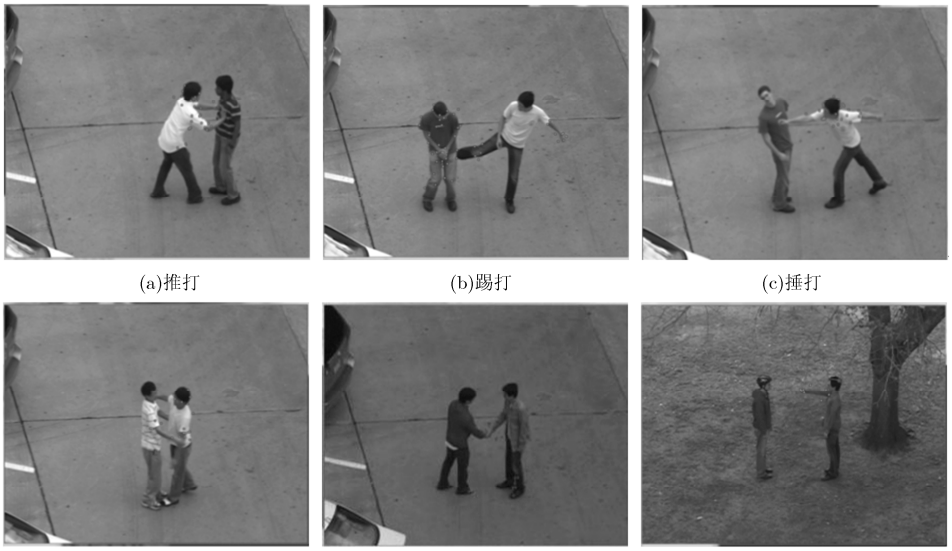
两人交互异常行为的识别是基于感兴趣帧，对感兴趣帧提取时空兴趣点，并采用增强单词包算法结合二分类 SVM 实现分类识别。图 10 是在 UT-interaction 数据库上检测到的特征点。

表 2 展示了在 UT-interaction 数据集上采用感兴趣帧技术前后各项指标的对比情况。

由表 2 可以看出，选取感兴趣帧前后，基于增强单词包算法的交互异常行为检测都取得了较好的识别效果。选取感兴趣帧之后，不仅降低了计算机的存储空间，大幅度降低了算法耗时，提高了运算

表 2 选取感兴趣帧前后对比情况

比较	没有选取感兴趣帧	选取感兴趣帧	对比结果
特征点检测及描述子计算的时间(s)	52062	31680	下降 39.1%
增强单词包形成的时间(s)	15840	6840	下降 56.8%
异常行为分类精度(%)	93.3	95.0	提高 1.7%



(a)推打　(b)踢打　(c)捶打

(d)拥抱　(e)握手　(f)用手指人

图 10　UT-interaction 数据库特征点检测结果展示

效率。而且避免了对无关帧的计算，减少了误判的兴趣点，分类准确度也有了一定程度提高。

5　结束语

本文提出了一种融合兴趣点的表观特征和时空位置分布特征的增强单词包算法用于行为识别。在单人行为数据集 KTH 上进行实验，并与传统的单词包算法进行对比，验证了增强单词包算法的优越性。在多人交互行为数据库 UT-interaction 上提出筛选感兴趣帧技术，不仅减少了算法的耗时，而且提高了分类精度。在下一步的工作中，我们计划将在更加复杂真实的环境中验证算法的有效性。

参 考 文 献

[1] 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. doi: 10.3724/SP.J.1016.2013.02512.
HU Qiong, QIN Lei, and HUANG Qingming. Human action recognition review based on computer vision[J]. *Journal of Computer*, 2013, 36(12): 2512–2524. doi: 10.3724/SP.J.1016.2013.02512.

[2] BEBAR A A and HEMAYED E E. Comparative study for feature detector in human activity recognition[C]. IEEE the 9th International conference on Computer Engineering Conference, Giza, 2013: 19–24. doi: 10.1109/ICENCO.2013.6736470.

[3] LI F and DU J X. Local spatio-temporal interest point detection for human action recognition[C]. IEEE the 5th International Conference on Advanced Computational Intelligence, Nanjing, 2012: 579–582. doi: 10.1109/ICACI.2012.6463231.

[4] ONOFRI L, SODA P, and IANNELLO G. Multiple subsequence combination in human action recognition[J]. *IEEE Journal on Computer Vision*, 2014, 8(1): 26–34. doi: 10.1049/iet-cvi.2013.0015.

[5] FOGGIA P, PERCANNELLA G, SAGGESE A, *et al.* Recognizing human actions by a bag of visual words[C]. IEEE International Conference on Systems, Man, and Cybernetics, Manchester, 2013: 2910–2915. doi: 10.1109/SMC.2013.496.

[6] ZHANG X, MIAO Z J, and WAN L. Human action categories using motion descriptors[C]. IEEE 19th International Conference on Image Processing, Orlando, FL, 2012: 1381–1384. doi: 10.1109/ICIP.2012.6467126.

[7] LI Y and KUAI Y H. Action recognition based on spatio-temporal interest point[C]. IEEE the 5th International

[8] Conference on Biomedical Engineering and Informatics, Chongqing, 2012: 181–185. doi: 10.1109/BMEI.2012.6512972.

[9] REN H and MOSELUND T B. Action recognition using salient neighboring histograms[C]. IEEE the 20th International Conference on Image Processing, Melbourne, VIC, 2013: 2807–2811. doi: 10.1109/ICIP.2013.6738578.

[10] COZAR J R, GONZALEZ-LINARES J M, GUIL N, *et al*. Visual words selection for human action classification[C]. International Conference on High Performance Computing and Simulation, Madrid, 2012: 188–194. doi: 10.1109/HPCSim.2012.6266910.

[11] WANG H R, YUAN C F, HU W M, *et al*. Action recognition using nonnegative action component representation and sparse basis selection[J]. *IEEE Transactions on Image Processing*, 2014, 23(2): 570–581. doi: 10.1109/TIP.2013.2292550.

[12] BILINSKI P and BREMOND F. Contextual statistics of space-time ordered features for human action recognition[C]. IEEE the 9th International Conference on Advanced Video and Signal-based Surveillance, Beijing, 2012: 228–233. doi: 10.1109/AVSS.2012.29.

[13] ZHANG L, ZHEN X T, and Shao L. High order co-occurrence of visualwords for action recognition[C]. IEEE the 19th International Conference on Image Processing, Orlando, FL, 2012: 757–760. doi: 10.1109/ICIP.2012.6466970.

[14] SHAN Y H, ZHANG Z, ZHANG J, *et al*. Interest point selection with spatio-temporal context for realistic action recognition[C]. IEEE the 9th International Conference on Advanced Video and Signal-based Surveillance, Beijing, 2012: 94–99. doi: 10.1109/AVSS.2012.43.

[15] TIAN Y and RUAN Q Q. Weight and context method for action recognition using histogram Intersection[C]. The 5th IET International Conference on Wireless, Mobile and Multimedia Networks, Beijing, 2013: 229–233. doi: 10.1049/cp.2013.2414.

[16] LAPTEV I and LIDEBERG T. Space-time interest points[C]. IEEE the 9th International Conference on Computer Vision, Nice, France, 2003: 432–439. doi: 10.1109/ICCV.2003.1238378.

[17] KLASER A, MARSZALEK M, and SCHMID C. A spatio-temporal descriptor based on 3D-gradients[C]. The 19th Conference on British Machine Vision and Pattern Recognition, Leeds, United Kingdom, 2008: 1–10.

张 良： 男，1970 年生，教授，主要研究方向为图像处理、模式识别、智能视频分析.