Journal of System Simulation

Volume 30 | Issue 7 Article 9

1-8-2019

Fusing Local and Global Features for Human Action Recognition

Tang Chao
1. Department of Computer Science and Technology, Hefei University, Hefei 230601, China;;

Miaohui Zhang
2. Energy Research Institute, Jiangxi Academy of Sciences, Nanchang 330096, China;;

Li Wei
3. School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 360054, China;;

Cao Feng
4. School of Computer and Information Science, Shanxi University, Taiyuan 030006, China;;

See next page for additional authors

Fusing Local and Global Features for Human Action Recognition

Abstract

Abstract: Recognizing human actions according to video features is an important research topic in a wide scope of applications. *In this paper, we propose a robust human motion detection method that combines canny operator with the combination of local and global optic flow methods. Meanwhile, this paper presents a simple but efficient action recognition algorithm using fusion visual features. The mixed features fuse two action descriptors, namely centre distance-based space time interest point and curvature function-based Fourier descriptors. The frame-based human action classifier is developed using random forests algorithm.* Experimental results show that the proposed method is accurate, efficient and robust compared with other supervised action recognition algorithms.

Keywords

human action recognition, local features, global features, space-time interest points, Fourier descriptors, random forest

Authors

Tang Chao, Miaohui Zhang, Li Wei, Cao Feng, Xiaofeng Wang, and Xiaohong Tong

Recommended Citation

Tang Chao, Zhang Miaohui, Li Wei, Cao Feng, Wang Xiaofeng, Tong Xiaohong. Fusing Local and Global Features for Human Action Recognition[J]. Journal of System Simulation, 2018, 30(7): 2497-2506.

第 30 卷第 7 期 系统仿真学报© Vol. 30 No. 7
2018 年 7 月 Journal of System Simulation Jul., 2018

融合局部与全局特征的人体动作识别

唐超[1]，张苗辉[2*]，李伟[3]，曹峰[4]，王晓峰[1]，童晓红[5]

(1. 合肥学院 计算机科学与技术系, 合肥 230601； 2. 江西省科学院 能源研究所, 南昌 330096； 3. 厦门理工学院 计算机与信息工程学院, 厦门 360054；4. 山西大学 计算机与信息技术学院, 太原 030006；5. 合肥职业技术学院 信息中心, 合肥 238000)

摘要：根据视频特征来识别人体行为是一个具有广泛应用的重要研究课题。*提出了一种鲁棒性强，抗噪性能优的人体运动目标检测方法和一种简单高效的多信息融合的混合行为特征表示方法和相应的识别算法。该混合行为特征具有简单、鲁棒和判别能力强的特点，它融合了基于中心距的时空兴趣点局部特征和基于曲率函数的傅里叶描述子全局特征，利用泛化能力较强的随机森林模型进行快速分类。*实验结果表明，该方法具有简单、快速和高效的特点。

关键词：人体行为识别；局部特征；全局特征；时空兴趣点；傅里叶描述子；随机森林

中图分类号：TP391 文献标识码：A 文章编号：1004-731X (2018) 07-2497-11

DOI: 10.16182/j.issn1004731x.joss.201807009

Fusing Local and Global Features for Human Action Recognition

Tang Chao[1], *Zhang Miaohui*[2*], *Li Wei*[3], *Cao Feng*[4], *Wang Xiaofeng*[1], *Tong Xiaohong*[5]

(1. Department of Computer Science and Technology, Hefei University, Hefei 230601, China; 2. Energy Research Institute, Jiangxi Academy of Sciences, Nanchang 330096, China; 3. School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 360054, China; 4. School of Computer and Information Science, Shanxi University, Taiyuan 030006, China; 5. Information Center, Hefei Technology College, Hefei 238000, China)

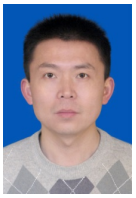
Abstract: Recognizing human actions according to video features is an important research topic in a wide scope of applications. *In this paper, we propose a robust human motion detection method that combines canny operator with the combination of local and global optic flow methods. Meanwhile, this paper presents a simple but efficient action recognition algorithm using fusion visual features. The mixed features fuse two action descriptors, namely centre distance-based space time interest point and curvature function-based Fourier descriptors. The frame-based human action classifier is developed using random forests algorithm.* Experimental results show that the proposed method is accurate, efficient and robust compared with other supervised action recognition algorithms.

Keywords: human action recognition; local features; global features; space-time interest points; Fourier descriptors; random forest

引言

基于视觉的人体运动分析是计算机视觉研究领域一个非常活跃的研究方向。它在虚拟现实、人机交互、基于内容的视频检索、智能视频监控和机器人等领域有着广泛的应用前景。人体运动分析技术包括：人体运动目标检测和分割、目标跟踪、行为识别和场景的语义理解。人体行为识别属于人的运动分析的高级处理部分。

人体行为识别可以看成一个时变数据的分类问题，包括行为特征提取和分类算法的设计两部



收稿日期：2017-08-14 修回日期：2018-01-05;
基金项目：国家自然科学基金(61672204, 41401521, 61602220), 安徽高校优秀拔尖人才培育资助项目(gxfx2017099);
作者简介：唐超(1977-)，男，安徽合肥，博士，讲师，研究方向为机器学习和计算机视觉。

分。基于视觉的人体行为识别的研究融合了图像处理、计算机视觉、模式识别和人工智能等学科的研究成果，是一个交叉的研究领域。

在复杂现实环境中，由于不受控因素的影响使得表征行为异常困难，一般只采用单一特征来表示行为，而且识别算法也是依靠长时间的视频数据才能完成识别过程。基于此，本文采用了多特征组合来表征行为，充分利用了不同特征的优势。同时识别算法采用多分类器组合策略的随机森林模型。

本文的主要贡献如下:

(1) 提出了一种基于局部与全局特征融合的人体动作识别方法，通过多视角的特征融合来表征人体行为，充分利用不同特征之间的优势来提升动作表征能力；

(2) 本文局部特征采用基于中心距的时空兴趣点特征，它不依赖底层的人体分割定位和跟踪，解决了对噪声和遮挡问题敏感的问题；而全局特征采用了剪影轮廓的基于曲率函数的傅里叶变换描述子，该特征包含非常多的人体信息，具有较强的动作表征能力，然后采用了特征级的特征融合策略，有效地提升人体动作表征能力；

(3) 采用随机森林进行建模分析，充分利用不同学习器学习偏置来提升学习器的泛化能力；

(4) 在公共数据集和自建的室内行为数据集上的实验结果表明本文提出的基于特征融合的人体行为识别算法取得较高的识别率。

1 相关工作介绍

目前有关行为识别的研究已经有许多的成果，有许多学者对此进行了综述与分析。其中，人体动作表征是人体动作识别过程中的关键。行为特征主要可以分为以下 5 种：基于形状的特征(Shape-based Features)、基于运动的特征(Motion-based Feature)、几何人体特征(Geometric Human Body Features)、兴趣点特征(Interest-point Features)、动态模型(Dynamic Model)和混合特征(Fusing Features)。接下来，我们概述已有的部分基于混合特征的人体行为识别工作。

Xin 等[1]提出了基于认识科学的数据还原方法和一种混合的“网络对网络”学习框架，通过静态、动态和序列混合特征来解决 3 个基本问题：动作空间域的变化，动作时间域的多变性，动作类内和类间差异，以此提高行为表征能力。Mohammadi 等[2]提出了一种基于集成支持向量机的人体行为识别方法，该方法采用 4 种不同的低层视觉特征 HOG(Histogram of Oriented Gradients), HOF (Histograms of Optical Flow), MBH(Motion Boundary Histograms)和轨迹(Trajectory)分别训练生成单个学习器，然后采用 DS (Dempster-Shafer) 策略对多个学习器输出结果进行融合。在传统的特征包模型(Bag-of-Features Model)中一般忽略了局部特征的空间关系，为了解决该问题 Yang 等[3]提出一种融合动作特征与多空间尺度组合(Multi-spatial-scale Configuration)的混合特征来表征人体动作取得较好的识别率。Liu 等[4]通过全局特征和局部特征结构变量和反映特征因果关系的混合特征描述符来实现人的行为识别，从而提高人的行为识别率。Li 等[5]提出了一种基于分层识别框架和 Boosting 算法的行为识别系统，该系统中将光流直方图，梯度方向直方图，Hu 矩，块剪影(Block Silhouettes)和自相关矩阵(Self-similarity Matrices)五种特征进行融合来表征行为，识别算法采用了基于 BP 神经网络的多类 AdaBoost 算法，并取得较好的识别效果。动作特征的选择直接影响到人体动作识别方法的效果。由于许多因素如人体的外观，环境和摄像机，往往会影响到单一的动作特征，因此，动作识别的准确性是有限的。Guo 等[6]在研究人类行为的表示和识别的基础上，充分考虑各种特征的优缺点，提出了一种结合全局轮廓特征和局部光流特征的混合特征，并取得较好的识别效果。

上述混合特征一般都是将运动特征如光流特征与形状统计特征进行简单融合，计算量大，抗噪性差，遇到遮挡的情况下，识别效果也不好。本文提出的基于局部与全局特征的人体动作识别方法，

DOI: 10.16182/j.issn1004731x.joss.201807009

将局部的时空兴趣点特征与全局的轮廓特征进行有效融合来表征人体动作，不仅解决了遮挡和噪声问题，在很大程度上保留了动作的绝大部分有效信息。局部特征与全局特征可有效进行互补，使动作的表征能力更强和有效。识别算法本文采了随机森林算法，可以有效提高识别率。

2 方法框架

基于不同的特征的优势，本文将两种不同的行为特征进行融合作为行为描述子，然后采用随机森林算法进行训练识别，系统流程见图 1。该方法保留了简单特征的高效计算便利，同时又保证了特征的鲁棒性和判别能力，而且该方法能够从两帧视频帧中快速完成行为识别。方法框架主要包括以下几个步骤：

(1) 给出标记了行为类别的训练视频集合，然后提取出运动人体感兴趣区域，为计算局部特征和全局特征做好准备；

(2) 从感兴趣区域中计算出基于中心距的时空兴趣点特征，然后再计算出基于曲率函数的傅里叶描述子特征，并分别对这两种特征进行规格化处理，然后采用一定的特征融合策略后作为人体行为的描述特征；

(3) 利用组合特征的训练样本集来训练随机森林模型；

(4) 识别时将预测图像进行相同的特征提取过程，然后输入到训练好的随机森林模型中进行识别，输出识别结果。

3 目标检测与特征提取

本节主要介绍人体感兴趣区域的提取过程，并详细给出局部特征与全局特征提取方法。其中，3.1 节给出了基于光流与边缘检测方法融合的运动目检测方法；3.2 节描述了基于中心距的时空兴趣点特征描述子；3.3 节介绍基于曲率函数的傅里叶描述子特征提取方法；在 3.4 节将给出多特征混合方法。

3.1 人体区域检测方法

基于视频的人体行为分析首要的和最重要的一步就是要进行运动人体的检测。本文提出了将 Canny 边缘检测算子和联合局部与全局约束的微分光流技术 CGOF(the Combining Local and Global Optic Flow Methods, CLGO)[7]相结合的新方法，通过 Canny 算子检测出物体的边缘，用 CLGO 微分光流法计算出运动目标的光流场信息，接着将目标的边缘形状信息和运动场信息进行整合，并采用数学形态学方法进行处理，从而得到最终的运动目标。具体做法如下：

Step1：图像的预处理。输入视频序列图像的相邻两帧图像，采用标准差为 1.5 的高斯滤波分别对图像进行预处理，去除图像中的噪声，$I(x,y)$ 是原始图像，$f(x,y)$ 是滤波后的图像。$K_{1.5}$ 是 $\sigma=1.5$ 的高斯滤波函数。

$$f(x,y)=K_{1.5}*I(x,y) \tag{1}$$

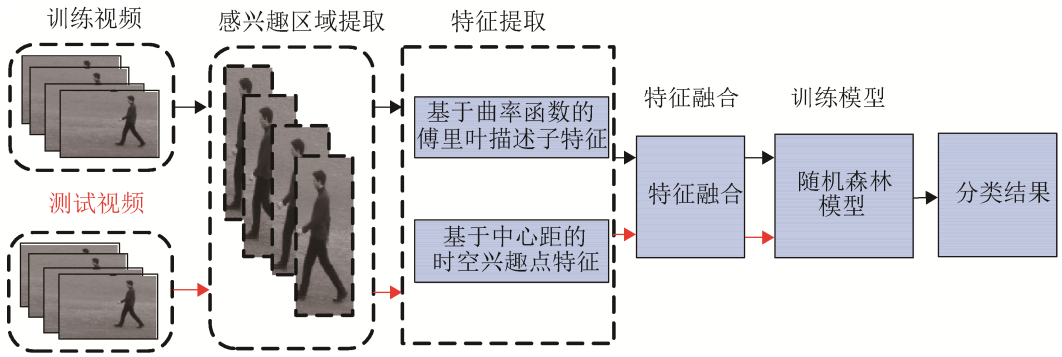


图 1 人体行为识别框架图

Fig.1 Framework of human action recognition

Step2：图像的边缘检测。使用 Canny 边缘检测算子获取图像中物体的边缘信息，得到边缘图像 $f_{EC}(x,y)$，它描述了图像的空间梯度信息。

Step3：图像的光流场计算。使用 CLGO 光流算法计算出整幅图像的光流场，得到光流场图像 $f_{CLGO}(x,y)$，它描述了图像的时间梯度信息。

Step4：图像的二值化。为了能动态地选取二值化分割阈值，采用最大类间方差法(OTSU 法)分别对边缘图像和光流场图像进行二值化，分别得到二值化边缘图像 $f_{BEC}(x,y)$ 和二值化的光流场图像 $f_{BCLGO}(x,y)$。

Step5：数据的融合。采用“与”运算对二值化后的边缘图像和光流场图像进行融合，得到融合后的目标图像。如公式(2)所示：

$$f_{FUSION}(x,y)=\begin{cases}1, & f_{BEC}(x,y)=1 \text{ and } f_{BCLGO}(x,y)=1\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

Step6：图像数学形态学处理。由于融合后的图像 $f_{FUSION}(x,y)$ 可能存在一些小面积区域是非运动目标，则需要对其进行腐蚀(Erosion)、膨胀(Dilation)、开运算(Open Operation)和闭运算(Close Operation)等操作。

Step7：图像区域填充和连通区域标记处理。经过腐蚀、膨胀、开运算、闭算法等数学形态学处理后，运动目标大部分被检测出来了，为了去除运动目标内可能存在的“空洞”，需要进一步进行区域填充(Filling Processing，FP)和连通区域标记(Connected Components Labeling，CCL)处理，最后可以提取出运动目标图像 Object(x,y)，结果如图 2 所示。

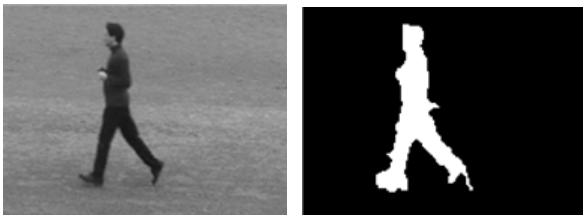


图 2 运动目标检测
Fig. 2 Moving object detection

3.2 基于中心距的时空兴趣点特征

人体行为识别局部特征提取是指提取运动人体中感兴趣的点或者块。这些点通常是视频运动中发生突变的点，而这些点包含了人体运动的大部分信息。局部特征对人体的表观变化，视角变化和部分遮挡问题具有较强的鲁棒性。时空兴趣点是一种典型的局 部特征。兴趣点的检测通常包含在空间域和在时间域的检测。

Laptev 将 2D 的 Harris 角点[8]扩展到 3D 的 Harris 角点[9]并作为时空域中显著变化的点。首先建立视频序列的线性空间表示：

$$L(:,\sigma_l^2,\tau_l^2)=g(:,\sigma_l^2,\tau_l^2)\times f(\cdot) \tag{3}$$

则可建立矩阵:

$$\boldsymbol{N}=g(:,\sigma_l^2,\tau_l^2)*\begin{vmatrix}L_x^2 & L_xL_y & L_xL_t\\ L_xL_y & L_y^2 & L_yL_t\\ L_xL & L_yL_t & L_t^2\end{vmatrix} \tag{4}$$

式中：$g(:,\sigma_l^2,\tau_l^2)$ 是高斯核函数；σ_l^2 是空间尺度因子；τ_l^2 是时间尺度因子；L 是经过平滑处理后的视频序列；矩阵 $\boldsymbol{N}$ 的 3 个特征值λ_1，λ_2 和λ_3 分别对应视频序列 L 在空间域(x,y)和时间域 t 三个方向的变化。当这 3 个值都比较大的时候，表示视频在这 3 个方向上变化都很大，那么这一点也即是时空兴趣点。

Laptev 将兴趣点的响应函数定义为：

$$H=\det(N)-k\times\operatorname{trace}^3(N)=\lambda_1\lambda_2\lambda_3-k(\lambda_1+\lambda_2+\lambda_3)^3 \tag{5}$$

式中：k 为系数，通常取值为 0.005。

Dollar[10]指出 Laptev 方法存在一个缺点，即检测出来稳定的兴趣点的数量太少，因此 Dollar 单独的在时间维和空间维先采用 Gabor 滤波器进行滤波，这样的话检测出来兴趣点的数目就会随着时间和空间的局部邻域尺寸的改变而改变。本文采用 Dollar 方法检测时空兴趣点，定义响应函数为：

$$R=(f\times g\times h_{ev})^2+(f\times g\times h_{od})^2 \tag{6}$$

$$h_{ev}(t;\tau,\omega)=-\cos(2\pi t\omega)e^{\frac{-t^2}{\tau^2}} \tag{7}$$

$$h_{od}(t;\tau,\omega)=-\sin(2\pi t\omega)e^{\frac{-t^2}{\tau^2}} \tag{8}$$

式中：g 是高斯平滑核函数；h_{ev} 和 h_{od} 是作用于时间维度上的一对正交 Gabor 滤波器。通常设定为参数 $\omega=4/\tau$

为了保证时空兴趣点集的平移和缩放性，本文设计了一种基于中心距的时空兴趣点归一化方法。假设 $P=\{p_1,p_2,...,p_n\}$ 是视频序列 $f(x,y,t)$ 某一时刻 t 对应的时空兴趣点集，n 为采集到的时空兴趣点个数，每一个兴趣点可由 $p_i(x_i,y_i)$ 来表示。

本文采用人体中心点作为极点 O。这样兴趣点 $p_i(x_i,y_i)$ 在图像坐标系下的坐标就可以转换为以人体中心点 O 为极点的极坐标表示形式，如式(9)所示：

$$(r_i,\theta_i)=\begin{cases} r_i=\sqrt{(y_i-y_c)^2+(x_i-x_c)^2} \\ \theta_i=\arctan\left(\dfrac{y_i-y_c}{x_i-x_c}\right) \end{cases} \quad i=1,2,...,n \tag{9}$$

式中：(x_c,y_c) 和 (x_i,y_i) 分别是人体中心点和兴趣点在图像直角坐标系下的坐标。在得到兴趣点极坐标后，需要进一步对其进行归一化处理，为了保存数据落在区间[0, 1]之间，采用极小极大归一法，具体如公式(10)和(11)所示：

$$r_i'=\frac{r_i-\min\_r}{\max\_r-\min\_r} \tag{10}$$

$$\theta_i'=\frac{\theta_i-\min\_\theta}{\max\_\theta-\min\_\theta} \tag{11}$$

式中：$\min\_r$ 和 $\max\_r$ 分别是原始 r 中的最小值和最大值；$\min\_\theta$ 和 $\max\_\theta$ 分别是原始 θ 中的最小值和最大值。经过规一化处理后，就可以得到具有平移变换，尺度变换和旋转变换不变性的基于中心距的时空兴趣点行为描述子(Space Time Interest Point based on Center Distance, STIP-CD)，图 3 为检测到的兴趣点。

3.3 基于曲率函数的傅里叶描述子

最早的傅里叶描述子得益 Zahn 等[11]的工作，他把强有力的傅里叶理论应用于形状描述的。其主要思想是利用表示形状全频率分量的一组数字来描述轮廓特征。

图 3 时空兴趣点检测结果
Fig. 3 Detection result of spatial-temporal interest point

由于生成基于复坐标函数的傅里叶描述子过程中需要大量的复数计算过程，因此算法的时间复杂度非常高，不利于快速提取人体行为特征描述子。针对这一问题，本文采用了基于曲率函数(Curvature Function)的傅里叶描述子。首先通过 CLGO 算法检测得到运动人体目标并对其二值处理。经过二值化后的人体图像序列还需要经过一系列数学形态学的处理，以保证获得完整有效的人体轮廓曲线。

首先在人体轮廓线上等间隔采样取 100 个点，这些点的坐标为 $(x_i,y_i),i=1,2,...,N$，其中 N=100 为轮廓上取的离散点个数，具体如图 4 所示。

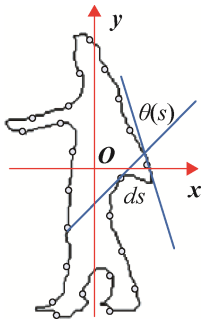


图 4 基于曲率函数的傅里叶描述子
Fig. 4 Fourier descriptor based on curvature function

廓轮线上某一点的曲率定义为轮廓切向角度相对于弧长的变化率。曲率函数 $K(s)$ 可表示为：

$$K(s)=\frac{d}{ds}\theta(s) \tag{12}$$

式中：$\theta(s)$ 是轮廓线的切向角度，定义为：

$$\begin{cases} \theta(s)=\arctan\left(\dfrac{y'_s}{x'_s}\right) \\ y'_s=\dfrac{dy}{ds} \\ x'_s=\dfrac{dx}{ds} \end{cases} \tag{13}$$

经过傅里叶变换后，空域中的曲率函数可以利用频域中的傅里叶变换系数来描述。对于曲率函数，仅考虑正频率的坐标轴，因为傅里叶变换系数 F_i 是对称的，即 $|F_i|=|-F_i|$。则基于曲率函数的形状描述子可表示为：

$$CFD_k=\{|CFD_1|,|CFD_2|,...,|CFD_{N/2}|\} \tag{14}$$

在得到曲率函数形状描述子后需要进一步对其归一化处理。采用对数函数转换法对其进行归一化，具体如式(15)所示：

$$CFD'_k=\frac{\log_{10}(CFD_k)}{\log_{10}(\max\_CFD_k)} \tag{15}$$

式中：$\max\_CFD_k$ 是 CFD_k 的最大值。图 5 为某一帧行为轮廓经过傅里叶变换后的系数，该 CFD_k 系数具有平移、旋转、尺度不变性，以及与曲线的起始点无关的特性。

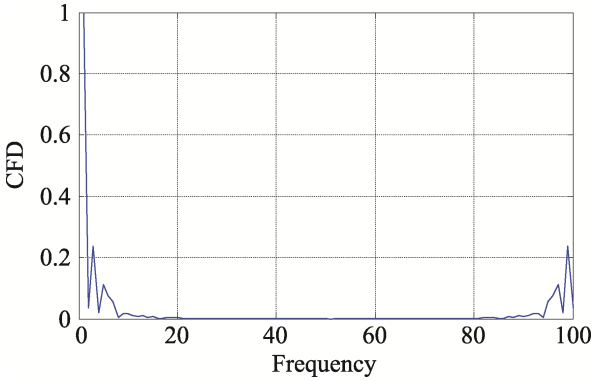


图 5 规格化后的基于曲率函数的傅里叶描述子
Fig. 5 Normalized Fourier descriptor based on curvature function

所以只记录傅里叶变换后能量集中的低频部分作为行为描述特征，即取前面 30 个低频分量作为描述人体轮廓特征向量，则可以得到基于曲率函数的傅里叶描述子(Curvature Function-Based Fourier Descriptors, CFD)。

3.4 特征混合

现有的特征融合方法主要有两种。一种是将两组特征向量组合成一个联合向量，然后在高维实向量空间中提取特征。另一种是利用复数向量组合两组特征向量，然后在复向量空间中提取特征，这两种特征融合方法都能提高识别率。第一种方法一般称为串行特征融合方法(Serial Feature Fusion)，第二种方法也可称为并行特征融合方法(Parallel Feature Fusion)[12-13]。本文采用计算简单的串行特征融合方法，具体地本文将基于中心距的时空兴趣点特征与基于曲率函数的傅里叶描述子特征进行融合，将两种不同的特征组合在一起作为行为描述特征，即：

$$F_T=\{STIP-CD,CFD\} \tag{16}$$

4 识别方法

在本节中，详细介绍随机森林算法，并应用该算法进行建模识别过程。

4.1 随机森林

2001 年，加州大学伯克利分校的 Breiman[14] 提出了随机森林模型，如图 6 所示，随机森林是由若干树型分类器 $\{T_1,T_2,...,T_B\}$ 组成的集成分类器。随机森林使用决策树装袋，通过随机地从原训练集中有放回的地选取 N 个样本，将随机性加入到构建模型的过程中。下面，我们将重点介绍一下随机森林的训练过程以及它的一些特性。

首先作如下定义：

$\{T_1(X),T_2(X),...,T_B(X)\}$：是 B 个树型分类器的集合。

$X=\{x_1,x_2,...,x_p\}$：是 p 维的特征向量数据。

$\{\hat{Y}_1(X),\hat{Y}_2(X),...,\hat{Y}_B(X)\}$：是 B 个树的集成输出结果，其中，$\hat{Y}_B$ 是第 B 个树的分类预测结果。

$\hat{Y}$：是 B 个树的集成输出结果。对于分类问题而言，是 B 个树的多数投票结果预测；对于回归问题，则是 B 棵树的实值预测结果的平均。

DOI: 10.16182/j.issn1004731x.joss.201807009

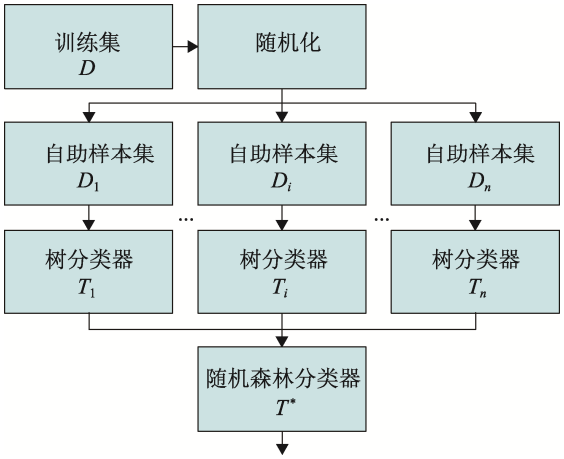


图 6 随机森林示意图
Fig. 6 Random forest algorithm diagram

具体的随机森林算法为：

随机森林算法

输入：

N个训练数据样本集合：

$D=\{(X_1,Y_1),(X_2,Y_2),...,(X_N,Y_N)\}$；

其中 $X_i, i=1,2,...,N$ 是样本的特征向量，Y_i 是对应的类别。

过程：

1. 从训练样本中，自举抽取出n个训练样本 (随机有放回的采样)。

2. 对于每一个自举样本集合，用下列方式生成一棵树：在树的每一个子节点处，用抽取出的 m_{try} 维样本(不是整个特征维样本)生成树的子节点分支。直到树长大到不能再进行分支和进行剪枝为止。

3. 重复上述过程，直到生成B棵树为止。

当 $m_{try}=p$ 时，随机森林算法变转变成Bagging算法了。当样本的特征维度非常大的时候，随机森林算法是非常有效的，这是因为：第一，通常我们在生成一棵树的时候，是用所有的特征维度来生成节点并进行分支，而随机森林只用 $m_{try}<<p$ 个维度数据来生成节点，因此它的搜索速度更快；第二，为了得到最优的模型，对于单棵决策树而言，需要进行剪枝操作，这通常是经过交叉验证(Cross-validation)来完成的，因此会占用大量的计算时间，而随机森林不需要进行剪枝操作。我们发现，当数据的特征维度特别巨大的时候，随机森林可以用比普通单棵决策树更少的时间完成训练过程。

4.2 识别算法

基于单帧的随机森林行为识别过程如下：一个行为视频帧 I 通过每个决策树分类器 $\{T_1,T_2,...,T_N\}$ 输出 M 个置信度 $p(,I)|_{(f(I)=c)}$，每个置信度表示该视频帧 I 属于第 c 类行为的概率，$c\in\{1,2,...,M\}$；最后，随机森林的决策结果是基于所有决策树结果的平均，如式(17)所示：

$$\tilde{F}(I)=\text{argmax}_c\left\{\frac{1}{N}\sum_{n=1}^{N}p(n,I)|_{(f(I)=c)}\right\} \tag{17}$$

随机森林属于 Bagging 类型模型，训练过程与 Bagging 类似，在样本的选择上避免了模型的过拟合(Over Fitting)问题，还在 Bagging 的基础上引入属性随机选择，从而保证了随机森林的分类性能。

5 实验与结果

本节将介绍在自建和公共的数据集上实验结果和分析。

5.1 度量标准

实验中，采用了交叉验证的方法(Cross-Validation)来训练识别模型和测试性能。同时采用精确度(Precision)、召回率(Recall)和 F 值(F-measure)作评价手段来衡量算法的效果，如式(18)~(20)所示：

$$P=TP/(TP+FP) \tag{18}$$

$$R=TP/(TP+FN) \tag{19}$$

$$F=2BR/(R+P) \tag{20}$$

对于二分类问题来说，TP 对应于被分类模型正确预测的正样本数；FP 对应于被分类模型错误预测为正类的负样本数；FN 为对应于被分类模型错误预测为负类的正样本数。这些公式可以推广到多分类问题中。

5.2 数据集

本文在Weizmann和KTH 2个公共数据集和自建的室人行为数据集上测试我们的方法，从数据集大小和数据特点等方面分别介绍这 3 个数据集。

Weizmann[15]数据库包含 10 个动作分别是“running”, “walking”, “skipping”, “jumping-jacks”,

“jump forwards”, “jump in place”, “sideways gallop”, “two-handed wave”,“one-handed wave”和 “bend”, 每个动作有 10 个人执行。在这个视频集中，其背景是静止的，且前景提供了剪影信息。该数据集较为简单。KTH[16]行人数据库包含了 6 种动作，分别为“walking”, “jogging”, “running”, “boxing”, “hand waving”, 和 “hand clapping”。每种动作由 25 个不同的人完成。每个人在完成这些动作时又是在 4 个不同的场景中完成的，4 个场景分别为室外，室内，室外放大，室外且穿不同颜色的衣服。

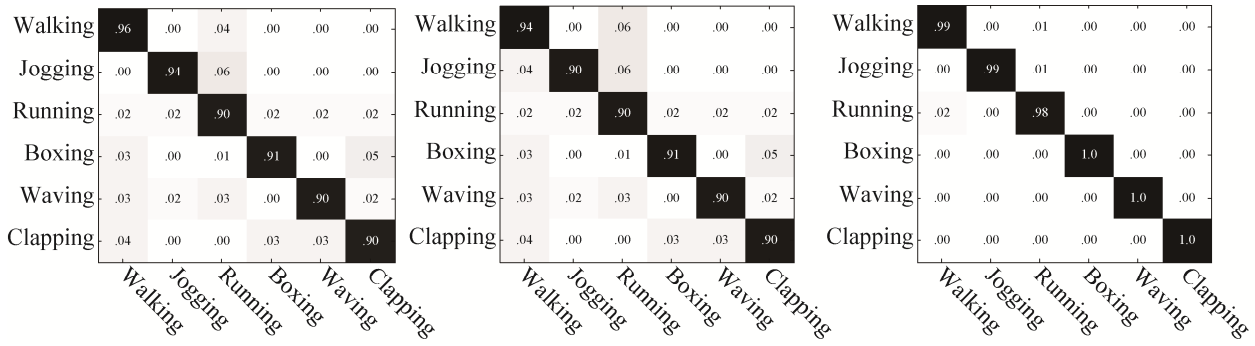
此外，我们通过 Kinect 传感器构建了一个室内行为数据库。微软 Kinect 的 RGB-D 传感器是专为微软 X-BOX 视频游戏而设计的控制设备的外设装置。通过 Kinect 传感器采集了十种不同的单一室内动作，分别是：“walk”、 “sit”、 “side walk”、“run”、 “pick-up”、 “jump”、 “hand wave”、 “hand clap”、“box”和“bend”。为了能更有效的提取出人体轮廓，使用了单一背景。每种行为帧数为 26~40 帧范围，图像分辨率为 320×240。

5.3 实验和结果

在本节中，我们进行 3 个实验来验证方法的可行性和高效性。第 1 个实验分别在 3 个不同的数据集上测试单一特征和采用混合特征后的识别率。第 2 个实验是分别在 3 个不同的数据集下的 Precision，Recall 和 F-Measure 情况。第 3 个实验是不同数据集下本文方法与其它算法对比情况。

实验 1 我们以混淆矩阵的方式给出本文方法的识别结果，图 7~9 以混淆矩阵的方式展示了本文提出的方法分别在基于中心距的时空兴趣点特征、基于曲率函数的傅里叶描述子特征和基于混合特征上的行为识别正确率结果。混淆矩阵的第(i, j)个元素表示第 i 类行为被分类为第 j 类行为的比例，因此对角线上的值越大，分类效果就越好。从上述图中可以看到当采用混合特征时，10 个动作类别的识别准确率都要高于前面的单一特征。

从图 7(a)中可以看出，在 KHT 数据集上，当采用时空兴趣点单一特征时，6 种动作的识别率普通不高，由于 walking，running 和 jogging 这 3 种动作的相似度较高，所以采用单一特征时 walking，running 和 jogging 这 3 种动作容易分错混淆。boxing，waving 和 clapping 这 3 种动作的误分类也较高，整体识别效果一般。从图 7(b)中可以看出，在采用基于曲率函数的傅里叶描述子特征时，walking，running 和 jogging 这 3 种动作误识率也较高，识别的时候易混淆，而 boxing，waving 和 clapping 这三种动作也同样识别率不高。从图 7(c)中可以看出，将二者结合的混合特征则很好地避免了这点，对 walking，running 和 jogging 的区分都能得到很理想的识别效果。boxing，waving 和 clapping 三个动作的识别率达到了 100%。在 Weizmann 和 Kinect 数据集上也有相似的结果，从图 8 和图 9 中可以看到当采用混合特征时，都取得较单一特征更高的识别率。

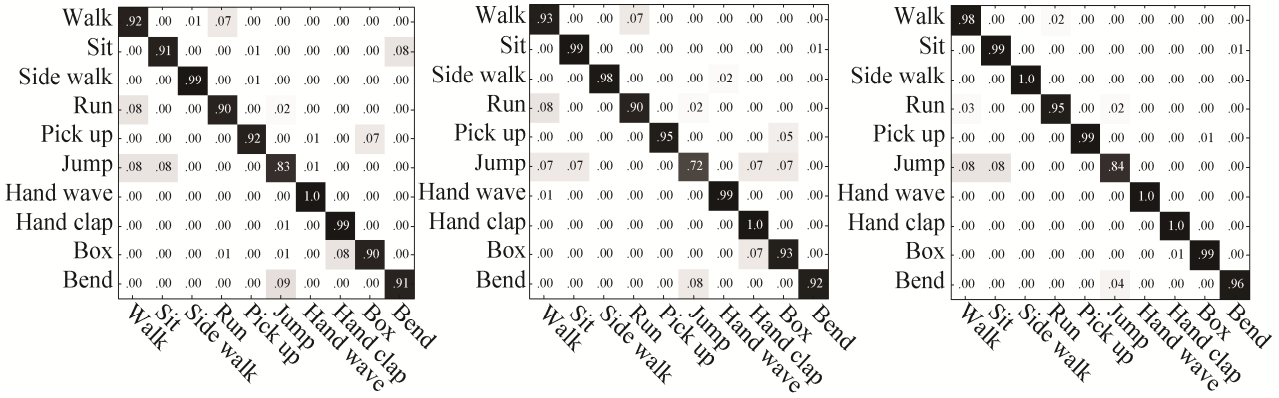


(a) 使用 STIP-CD 特征 (b) 使用 CFD 特征 (c) 使用混合特征

图 7 在 KHT 数据集上识别结果混淆矩阵

Fig. 7 Confusion matrix results on KHT dataset

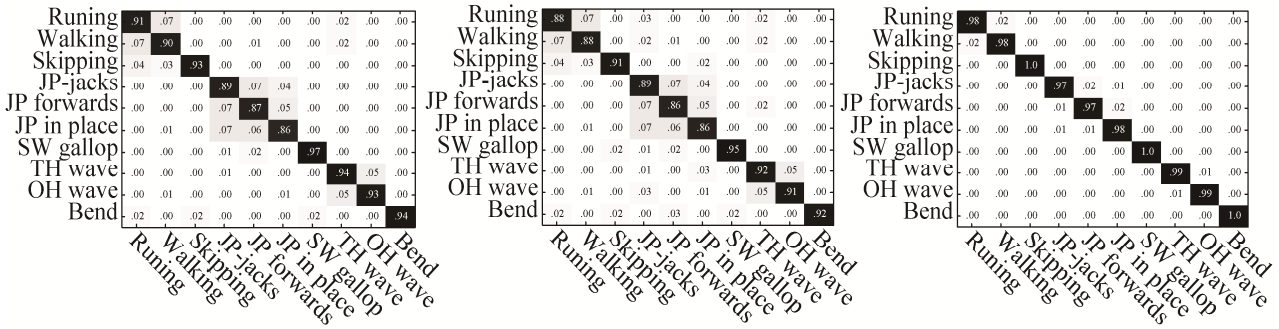
DOI: 10.16182/j.issn1004731x.joss.201807009



(a) 使用 STIP-CD 特征 (b) 使用 CFD 特征 (c) 使用混合特征

图 8 在 Kinect 数据集上识别结果混淆矩阵

Fig. 8 Confusion matrix results on Kinect dataset



(a) 使用 STIP-CD 特征 (b) 使用 CFD 特征 (c) 使用混合特征

图 9 在 Weizmann 数据集上识别结果混淆矩阵

Fig. 9 Confusion matrix results on Weizmann dataset

实验 2 给出了采用精确率、召回率和 F 值来评估不同特征及其组合情况下的行为识别率见表 1。从表 1 中可以看到采用 STIP-CD+CFD 组合特征的行为识别率高于采用单一特征 STIP-CD 和 CFD 特征表示下的行为识别率，可见多特征组合的行为表示有着较高的识别率。

在第 3 个实验中，我们将本文方法与其它方法进行了对比(如表 2 所示)。表 2 列举了本文的基于单视频帧的随机森林识别算法与最近邻(KNN)、Boosting、Bagging、支持向量机(SVM)和人工神经网络(ANNs)识别方法的比较。从表中的比较结果中，我们可以看出，基于 STIP-CD +CFD 特征组合的随机森林识别算法取得最高的识别率达到 97%。同时，随机森林识别算法整体性能也优于 KNN、Booting、Bagging 和 ANNs 方法，取得了与 SVM 相同的平均识别率，达到了 90%的正确率。

表 1 单特征与混合特征识别准确率

Tab. 1 Recognition accuracy of the proposed method using single type of descriptors and mixed descriptors /%

数据集	特征	Precision	Recall	F-Measure
KHT dataset	STIP-CD+CFD	97	97	97
	STIP-CD	96	96	96
	CFD	92	92	91
Kinect dataset	STIP-CD+CFD	96	96	96
	STIP-CD	86	86	86
	CFD	70	70	69
Weizmann dataset	STIP-CD+CFD	95	95	95
	STIP-CD	86	86	86
	CFD	70	70	69

表 2 不同行为识别方法比较

Tab. 2 Comparison of different action recognition methods /%

数据集	描述子	RF	KNN	Boosting	Bagging	SVM	ANNs
KHT dataset	STIP-CD+CFD	97	91	94	92	96	91
	STIP-CD	96	87	85	84	80	80
	CFD	92	85	88	84	80	74
Kinect dataset	STIP-CD+CFD	96	91	94	92	96	91
	STIP-CD	86	87	85	84	83	87
	CFD	70	65	70	71	65	60
Weizmann dataset	STIP-CD+CFD	95	91	94	92	96	91
	STIP-CD	86	80	85	80	84	83
	CFD	70	68	70	64	65	60

6 结论

本文提出了一种基于局部特征与全局特征融合行为识别方法，将基于中心距的时空兴趣点局部特征与基于曲率函数的傅里叶描述子全局特征进行融合作为行为描述特征，并使用随机森林学习框架作为训练识别工具，取得较好的学习效果。与传统单一行为特征相比，本文方法具有较高的识别率，实时性，鲁棒性好的优点。行为识别是一个充满挑战性的开放性课题，虽然本文方法在自建的数据集上取得了较好的实验结果，但仍有许多问题需要深入研究。为了获得泛化能力强的分类模型，通常需要大量的标记训练视频样本，这需要大量的人工标记劳力，这给建模带来实际的困难。如何利用大量触手可及的未标记视频样本来提升学习系统的性能成为一个值得研究的方向。

参考文献:

[1] XIN M, ZHANG H, WANG H, et al. Arch: Adaptive Recurrent-Convolutional Hybrid Networks for Long-Term Action Recognition [J]. Neurocomputing (S0925-2312), 2016, 178: 87-102.

[2] MOHAMMADI E, WU Q M J, SAIF M. Human Activity Recognition Using an Ensemble of Support Vector Machines[C]//Proc of the 2016 International Conference on High Performance Computing & Simulation. NEW YORK, NY: IEEE, 2016: 549-554.

[3] YANG J-F, MA Z, XIE M. Multiscale Spatial Position Coding under Locality Constraint for Action Recognition [J]. Journal of Electrical Engineering & Technology (S1975-0102), 2015, 10(4): 1851-1863.

[4] LIU X, LI Y. Research on Human Action Recognition Based on Global and Local Mixed Features[C]// Proc of the 2014 International Conference on Mechatronics, Control and Electronic Engineering. NEW YORK, NY: IEEE, 2014, 113: 692-696.

[5] LI N, CHENG X, GUO H, et al. Human Action Recognition Based on Multi-Feature Fusion and Hierarchical BP-Ada Boost Algorithm [J]. Journal of Southeast University (Natural Science Edition) (S1003-7985), 2014, 44(3): 493-498.

[6] 郭利, 姬晓飞, 李平, 等. 基于混合特征的人体动作识别改进算法[J]. 计算机应用研究, 2013, 30(2): 601-604.
GUO L, JI X F, LI P, et al. Mixed Features Based Improved Human Action Recognition Algorithm [J]. Application Research of Computers, 2013, 30(2): 601-604.

[7] BRUHN A, WEICKERT J, SCHNORR C. Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods [J]. International Journal of Computer Vision (S0920-5691), 2005, 61(3): 1-21.

[8] HARRIS C. A Combined Corner and Edge Detector[C]// Proc of the 4th Alvey Vision conference. 1988: 147-151.

[9] LAPTEV I, LINDEBERG T. Space-Time Interest Points[C]// Proc of the 9th IEEE International Conference on Computer Vision. NEW YORK, NY: IEEE, 2003: 432-439.

[10] DOLLAR P, RABAUD V, COTTRELL G, et al. Behavior Recognition Via Sparse Spatio-Temporal Features[C]// Proc of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. NEW YORK, NY: IEEE, 2005: 65-72.

[11] ZAHN G T, ROSKIES R Z. Fourier Descriptors for Plane Closed Curves [J]. IEEE Transactions on Computers (S0018-9340), 1972, C-21(3): 269-281.

(下转第 2514 页)

https://dc-china-simulation.researchcommons.org/journal/vol30/iss7/9
DOI: 10.16182/j.issn1004731x.joss.201807009