# 结合稠密轨迹与视频显著性特征的人体动作识别

高德勇[1,2]，康自兵[1*]，王松[1,2]，王阳萍[1,3]

[1]兰州交通大学电子与信息工程学院，甘肃 兰州 730070；

[2]甘肃省人工智能与图形图像工程研究中心，甘肃 兰州 730070；

[3]甘肃省轨道交通装备系统动力学与可靠性重点实验室，甘肃 兰州 730070

**摘要** 传统稠密轨迹算法在人体动作识别中取得了较大的成功，但是其在轨迹的形成过程中将动作产生的轨迹和背景运动导致的轨迹进行了相同处理，导致视频表示过于冗余，识别精度受限。为解决这一问题，首先分析背景运动与行为运动模式的差异性，以特征字典的稀疏系数矩阵为基础，利用低秩分解的方法得到稀疏误差矩阵，进一步求解出视频的显著图，然后以显著图作为依据仅在动作相关区域内形成显著性轨迹，并以此表征人体动作。最后基于公开数据集：UCF Sports 数据集和 YouTube 数据集，验证了本文方法的有效性。

**关键词** 图像处理；动作识别；稠密轨迹；视频显著性；低秩矩阵分解；稀疏编码

**中图分类号** TP391 **文献标志码** A **doi**：10.3788/LOP57.241003

# Human-Body Action Recognition Based on Dense Trajectories and Video Saliency

Gao Deyong[1,2], Kang Zibing[1*], Wang Song[1,2], Wang Yangping[1,3]

[1] *School of Electronic & Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu* 730070, *China*;

[2] *Gansu Provincial Engineering Research Center for Artificial Intelligence and Graphic & Image Processing, Lanzhou, Gansu* 730070, *China*;

[3] *Gansu Provincial Key Laboratory of System Dynamics and Reliability of Rail Transport Equipment, Lanzhou, Gansu* 730070, *China*

**Abstract** The traditional dense trajectory algorithm has achieved great success in human-body action recognition. However, the trajectories of the action and background motions are processed equally during algorithm′s formation, which leads to redundant video representation and limited recognition accuracy. In this paper, the patterns of the background and behavioral motions are compared, a sparse error matrix is obtained using low-rank matrix decomposition on the basis of the sparse coefficient matrix of the feature dictionary, and a saliency map is solved. The saliency map is then used as the base for representing human-body action in only the action-related areas. The validity of this method is confirmed based on the open datasets UCF Sports and YouTube.

**Key words** image processing; action recognition; dense trajectories; video saliency; low-rank matrix decomposition; sparse coding

**OCIS codes** 100.3008; 110.4153; 110.4155

## 1 引 言

人体动作识别是指通过机器视觉技术对摄像机采集到的视频图像数据进行有关的处理和分析，检测出图像序列中的行为动作，并通过相邻帧之间时空特征的关联性对其表征，最后对提取的特征进行

**收稿日期**：2020-03-30；**修回日期**：2020-04-19；**录用日期**：2020-05-29

**基金项目**：国家自然科学基金(61763025)、国家市场监督管理总局科技计划项目(2019MK150)、甘肃省科技计划项目(18JR3RA104)、甘肃省教育厅科技项目(2017D-08)

* **E-mail**：914764692@qq.com

学习并理解其中的行为动作。近年来随着人工智能相关技术的蓬勃发展，基于视频图像序列的人体动作识别已成为计算机视觉领域研究的热点问题，在视频检索、智能视频监控、人机交互、自动驾驶等领域得到了广泛的应用。

动作识别过程中如何提取有效且稳健的动作特征以及如何选取性能优良的分类器来对动作类别进行预测分类，对识别性能起着至关重要的作用[1]。由于人体动作在图像表观上类内以及类间变化差异较大，对于真实场景下的视频，动态背景的分析与处理是研究人员亟待解决的问题。背景运动一般由背景目标运动或者拍摄过程中相机运动引起，在整个视频中占据较大的范围甚至会覆盖人体动作，通常以相似的方式存在于视频中，这意味着背景运动一般在每个单独的帧中呈现统一的模式，背景运动的表示具有较大的相关性。相比之下，动作相关区域的运动是行为和背景运动的混合，在空间域中表现出不规则的特性。传统的动作特征表示方法中，基于全局特征[2]和局部特征[3-4]的动作特征表示已经日渐完善。Wang 等[5]提出了基于稠密轨迹(DT)的动作识别算法，该算法在诸多真实场景下的动作识别数据集上都取得了较好的识别结果。然而对于具有动态背景的视频，采样点落在了与动作无关的背景区域，产生了过多的背景轨迹，稠密轨迹算法在每一帧所有区域内的密集采样，并无区分动作区域和背景区域，在动作表征时若对动作相关轨迹和背景轨迹等同处理，必然会降低识别性能。为解决此类问题，Wang 等[6]利用前后帧图像存在的单应性关系，通过投影变换矩阵估计相机运动，消除了背景中由相机运动产生的干扰性轨迹，改进后的稠密轨迹算法(IDT)在鲁棒性和识别效果上取得了较大的提升。受人类视觉感知系统的启发，视觉显著性特征在静态图像和视频图像领域取得了较大的成功与广泛的应用[7-11]。Wang 等[12]针对视频中的特征表示，利用显著图来进行特征选择，进而稀疏提取信息丰富的兴趣点。Yi 等[13]首先计算原始稠密轨迹的表观显著性和运动显著性，然后对二者的显著性值利用线性组合的方式得到轨迹的联合显著性，去除了冗余轨迹，保留了与动作相关区域的轨迹。Somasundaram 等[14]在柯式复杂性和信息论的基础上，利用稀疏编码的残差确定显著的时空区域，并假设残差越大，时空块越显著。Li 等[15]通过运动显著性分析来提纯轨迹，并根据运动显著性的强度分布来优化词袋模型，用于从复杂视频的上下文中学习识别特征。Yi 等[16]利用显著性区域提取动作特征，降低了计算和存储的成本。本文受到视频显著性检测相关工作[17-18]上角标的启发，旨在真实场景的视频中获取与动作相关的区域，并在此区域中提取动作轨迹来表征人体动作。其中最关键的问题是在视频中找到动作相关区域，由于动作相关区域一般都具有较强的表观显著性和运动显著性，故可以将此任务简化为视频中行为显著区域的检测。

## 2 理论与方法

本文方法的框架如图 1 所示，图中 HOG 为方向梯度直方图，HOF 为光流直方图，MBH 为运动边界直方图。首先提取动作视频中的特征，对视频中的运动目标建立模型，构造运动信息的特征字典，然后对数据字典进行稀疏编码，得到稀疏系数矩阵。理论上，背景运动一般在视频帧中呈现出较高的相似性，其运动模式与行为动作相比更具有规则性和统一性[7-8]，所以表示背景运动的系数矩阵线性相关，由背景运动构成的特征矩阵具有低秩特性。然而，行为动作则无规律性，所以表示动作特征的矩阵是一个具有少量非零元素的稀疏矩阵。综上所述，最后将系数矩阵分解为构成背景运动的低秩矩阵和表示行为动作的稀疏误差矩阵，利用稀疏误差矩阵得到滤除背景运动的显著图。当得到整个视频的显著图时，将其融入到密集特征点的追踪过程中，使得稠密轨迹只从显著区域中提取。在密集跟踪的过程中，计算轨迹点的显著性值，通过比较轨迹点的显著性值与给定阈值来判断该点是否为有效轨迹点。本文依照原始稠密轨迹方法的计算流程，即计算所有轨迹的特征、编码轨迹特征，获取视频级表示，并使用支持向量机(SVM)预测动作类别标签。

### 2.1 显著性区域检测

本研究的重点是检测与动作相关的显著性区域，结合在视频图像领域中视觉显著性研究的相关工作[7-11]，将动作视频分解为稀疏表示的显著部分和均匀规则的非显著部分。假设视频 $\boldsymbol{V}_i=[I_1,I_2,\cdots,I_t,\cdots,I_T]$，$t\in[1,T]$，$I_t$ 表示第 $t$ 帧，将视频分割成不重叠的时空块，大小为 $s\times s\times t$，空域大小为 $s\times s$，时域长度为 $t$，以视频 $\boldsymbol{V}_i$ 为例，利用低秩分解的方法对其显著性展开论述。设 $\boldsymbol{F}$ 表示特征矩阵，并将 $\boldsymbol{F}$ 分解为低秩矩阵 $\boldsymbol{L}$ 和稀疏误差矩阵 $\boldsymbol{S}$，即

$$\boldsymbol{F}=\boldsymbol{L}+\boldsymbol{S}, \tag{1}$$

其中满足

$$\min_{\boldsymbol{L},\boldsymbol{S}}\operatorname{rank}(\boldsymbol{L})+\lambda\|\boldsymbol{S}\|_0,\ \text{s.t.}\ \boldsymbol{L}+\boldsymbol{S}=\boldsymbol{F}, \tag{2}$$

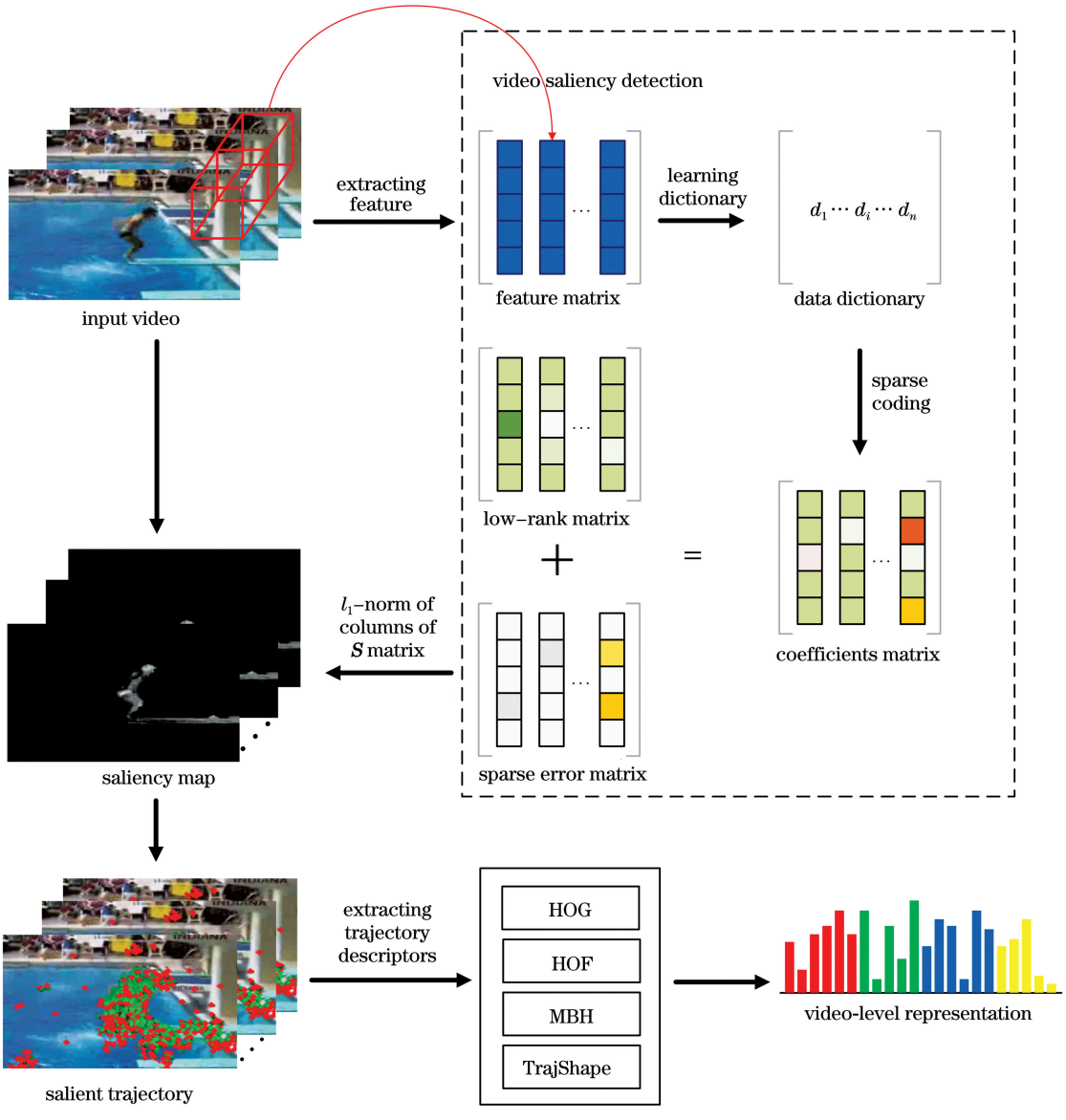


图 1 基于显著性的动作识别算法框架图

Fig. 1 Saliency-based action recognition algorithm framework

式中，rank($\boldsymbol{L}$)是低秩矩阵 $\boldsymbol{L}$ 的秩，$\lambda$ 是平衡低秩和稀疏性的权重因子。

由于(2)式是一个 NP-hard 问题，利用松弛凸优化方案，用 $l_1$ 范数替代 $l_0$ 范数，核范数替代 $\boldsymbol{L}$ 的秩，故(2)式可重新改写为

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1, \text{ s.t. } \boldsymbol{L}+\boldsymbol{S}=\boldsymbol{F}。 \quad (3)$$

在低秩分解前首先要找到一个合适的特征矩阵来作为低秩分解的输入，利用之前已划分视频 $\boldsymbol{V}_i$ 的时空块构造特征矩阵，即

$$\boldsymbol{X}=[\boldsymbol{X}_1 \quad \boldsymbol{X}_2 \quad \cdots \quad \boldsymbol{X}_i \quad \cdots \quad \boldsymbol{X}_n] \in \mathbf{R}^{m\times n}, \quad (4)$$

其中，$\boldsymbol{X}_i$ 是构成时空块 $i$ 的特征向量，$m$ 是向量的维数，$n$ 是向量的个数。对得到的特征矩阵进行字典学习，设字典 $\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_k]\in\mathbf{R}^{m\times k}$，则时空块的特征向量可以表示为

$$\boldsymbol{X}_i=\sum_{j=1}^{k}\boldsymbol{d}_j\beta_{ji}+\boldsymbol{\varepsilon}, \quad (5)$$

其中，$\boldsymbol{d}_j$ 是字典原子，$\beta_{ji}$ 是相对应的系数，$\boldsymbol{\varepsilon}$ 是高斯噪声组成的 $m$ 维向量。将(5)式改写为

$$\boldsymbol{X}_i=\boldsymbol{D}\boldsymbol{\beta}_i+\boldsymbol{\varepsilon}。 \quad (6)$$

因此，把 $\boldsymbol{X}_i$ 表示为稀疏编码 $\boldsymbol{\beta}_i=[\beta_{1i},\beta_{2i},\cdots,\beta_{ki}]\in\mathbf{R}^{k\times 1}$，也就是说每个特征向量都可以表示为字典中原子的稀疏线性组合。学习特征向量的稀疏表示可表示为

$$\min_{\boldsymbol{\beta}}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{\beta}\|_2^2+\lambda^*\|\boldsymbol{\beta}\|_1, \quad (7)$$

其中，$\boldsymbol{\beta}$ 是特征点稀疏表示的系数矩阵，$\lambda^*$ 是平衡稀疏性的正则化参数。在得到特征的稀疏表示后，使

用增广拉格朗日乘子(ALM)法通过优化(3)式来求解得到低秩矩阵和稀疏误差矩阵。求解稀疏误差矩阵 $\boldsymbol{S}$ 序列的 $l_1$ 范数来衡量对应时空块的显著性,最后得到视频序列的显著图。

## 2.2 稠密轨迹

基于轨迹的动作特征表示已经被证明是一种有效的特征表示方法。Wang 等[5]提出了更为有效的稠密轨迹来对动作进行描述,其算法原理如图 2 所示。

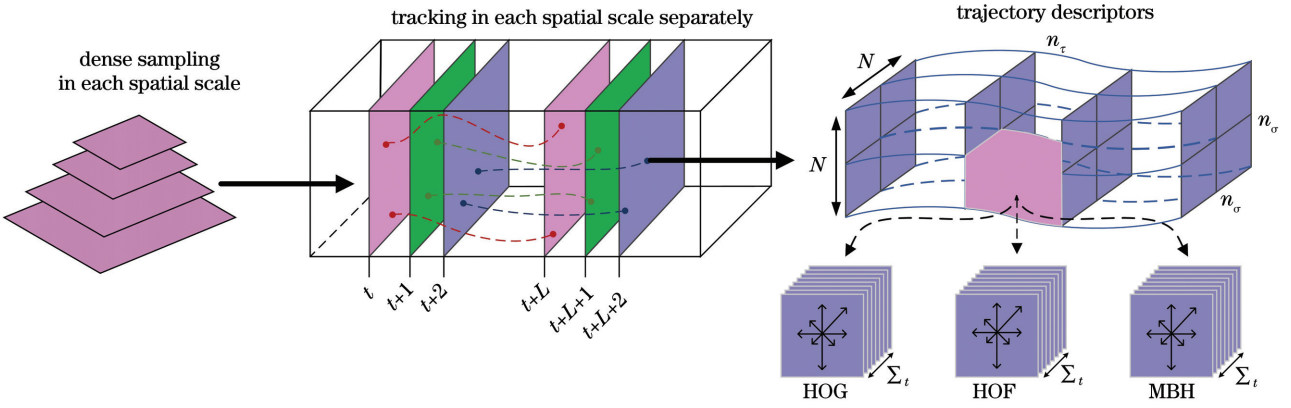


图 2 稠密轨迹算法原理图

Fig. 2 Dense trajectory algorithm framework

对于视频的每一帧,在 $w$ 个像素为间隔的网格上采样特征点并在不同的空间尺度下分别追踪采样点。在 $t$ 帧的特征点 $\boldsymbol{P}_t=(x_t,y_t)$ 通过在密集光流场 $\boldsymbol{\omega}_t=(u_t,v_t)$ 的中值滤波操作,可以在 $t+1$ 帧追踪得到,其位置定义为

$$\boldsymbol{P}_{t+1}=(x_{t+1},y_{t+1})=(x_t,y_t)+(\boldsymbol{M}*\boldsymbol{\omega}_t)\big|_{(x,y)}, \tag{8}$$

其中 $\boldsymbol{M}$ 是 3×3 的中值滤波核,$u_t$ 和 $v_t$ 分别为光流的垂直分量和水平分量。后续帧的特征点连接起来形成轨迹 $(\boldsymbol{P}_t,\boldsymbol{P}_{t+1},\boldsymbol{P}_{t+2},\cdots)$。轨迹的漂移是一个常见的问题,参照不同轨迹长度对识别精度的影响,采用默认的轨迹长度 $L=15$ frame,对轨迹的形状编码局部运动模式。给定一个长度为 $L$ 的轨迹,通过位移向量的序列 $\boldsymbol{S}$ 来描述轨迹的形状:

$$\boldsymbol{S}=(\Delta\boldsymbol{P}_t,\Delta\boldsymbol{P}_{t+1},\cdots,\Delta\boldsymbol{P}_{t+L-1}), \tag{9}$$

$$\Delta\boldsymbol{P}_t=(\boldsymbol{P}_{t+1}-\boldsymbol{P}_t)=(x_{t+1}-x_t,y_{t+1}-y_t), \tag{10}$$

然后将位移向量的大小求和,对轨迹的形状进行正则化表示,表达式为

$$\boldsymbol{S}'=\frac{(\Delta\boldsymbol{P}_t,\Delta\boldsymbol{P}_{t+1},\cdots,\Delta\boldsymbol{P}_{t+L-1})}{\sum_{j=t}^{t+L-1}\|\Delta\boldsymbol{P}_j\|}, \tag{11}$$

最终,将 $\boldsymbol{S}'$ 作为轨迹的唯一描述符。

## 2.3 基于显著性的轨迹

将显著图融入到密集跟踪的过程中,以提取视频中与动作相关的稠密轨迹,形成的轨迹中既具有重要的几何特性又具有较强的视觉显著性。在得到视频显著图后,假定动作区域内特征点的显著性强度值大于给定阈值 $T_s$,对任意的一条轨迹 $T_i$,计算它的每一个轨迹点 $P_{(t,i)}$ 的显著性值 $S(P_{t,i})$,然后与给定阈值 $T_s$ 比较,如果该点的显著性强度大于 $T_s$,则认为该点为有效的轨迹点,否则,在该点处终止轨迹 $T_i$ 的追踪。重新定义显著性轨迹,表达式为

$$T_{\text{salient}}=\{P_{t,i}\,|\,P_{t,i}\in T,\ S(P_{t,i})\geqslant T_s\}, \tag{12}$$

其中,$T$ 是稠密轨迹的集合。如果隶属于原始轨迹集合中任意特征点的显著性值大于给定的显著性阈值 $T_s$,则将该点归属于集合 $T_{\text{salient}}$。将 $T_{\text{salient}}$ 中的点代入(8)~(11)式重新确定轨迹的形状和位移向量。在整个与动作相关的显著性区域内重复以上过程,实现了仅保留与行为有关的轨迹、去除由背景运动导致的冗余轨迹。

## 2.4 动作分类

对于每一条显著性轨迹,计算 4 种类型的行为描述符:方向梯度直方图、光流直方图、运动边界直方图和轨迹形状(TrajShape),然后使用 Fisher 向量对这 4 种描述符进行统一编码,得到最终的视频表示,并将其输入到支持向量机(SVM)中预测动作类别。选取 $\chi^2$ 核函数的非线性支持向量机对动作进行分类:

$$K(H_p,H_q)=\exp\left[-\frac{1}{Z}D(H_p,H_q)\right], \tag{13}$$

其中 $H_p$ 和 $H_q$ 为两个视频的直方图,$D(H_p,H_q)$ 是两个直方图的 $\chi^2$ 距离,$Z$ 为所有训练样本直方图 $\chi^2$ 距离的平均值。

## 3 分析与讨论

### 3.1 实验数据集和实验环境

为了验证本文动作识别方法的有效性，在 UCF Sports 和 YouTube 两个公开的动作数据集上进行实验测试。

UCF Sports 数据集[19]总共有 150 个行为视频，分辨率为 720 pixel ×480 pixel，包含 10 种不同类的动作：跳水、打高尔夫、踢足球、举重、骑马、跑步、滑板、鞍马、单双杠和步行。该数据集具有场景大、视角广的特征，包含了大量的相机运动。图 3(a)给出 UCF Sports 数据集中的部分动作实例样本帧。

YouTube 数据集[20]中的视频来源于 YouTube 视频网站，是一个更具有挑战性的数据集。数据集总共有 1168 个视频序列，包含 11 个动作类：投篮、骑自行车、跳水、打高尔夫球、骑马、颠足球、荡秋千、打网球、蹦床、打排球和遛狗。YouTube 数据集包含运动相机和固定相机的混合、目标的尺度变化、视角和光线的变化、分辨率低、复杂背景和背景目标运动等干扰性因素。图 3(b)给出 YouTube 数据集中的部分动作实例样本帧。

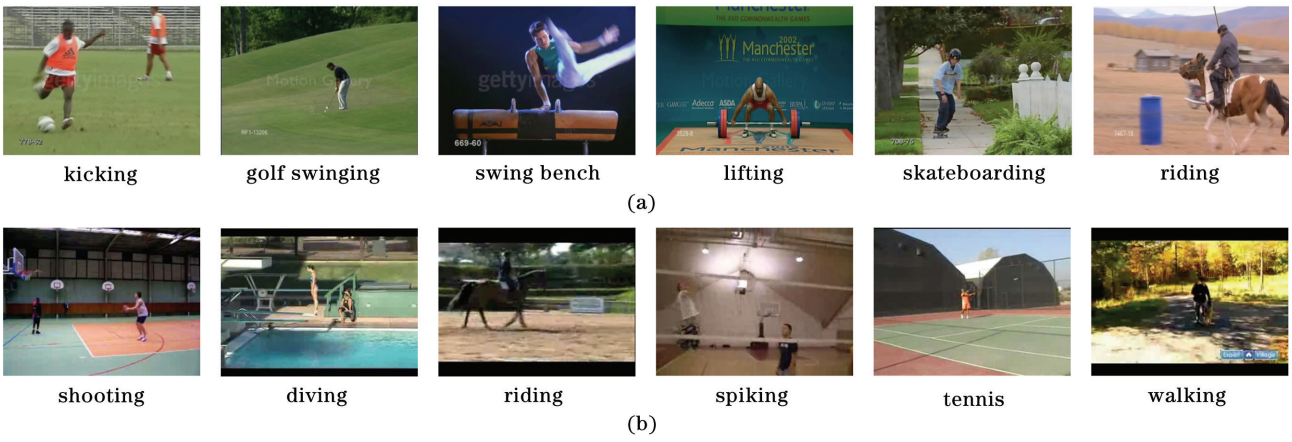


图 3 动作样本帧实例(UCF Sports/YouTube)。(a)UCF Sports 数据集；(b)YouTube 数据集

Fig. 3 Sample frames from UCF Sports and YouTube. (a) UCF Sports; (b) YouTube

本文方法以 Linux 操作系统为实验平台，采用 opencv2.4.9 作为开发环境。算法实现的环境如表 1 所示。

表 1 实验环境

Table 1 Experimental environment

| Experimental environment | Detail information |
|---|---|
| OS | Ubuntu14.04 |
| CPU | Intel(R) i7-8700 @ 3.20 GHz |
| GPU | Nvidia GeForce GTX 1060 3 GB |
| RAM | 16 GB |
| Compiler | Matlab2016 |

### 3.2 实验参数优化

为了降低计算代价，将数据集中的视频大小调整为原始视频的四分之一，用于显著性区域的检测。本研究中将视频分割为不重叠的时空块，构造特征矩阵，进一步将其作为显著性区域检测过程中低秩分解的输入，通过与给定阈值 $T_s$ 比较得到视频显著图。

在两个数据集上测试显著性阈值 $T_s$ 对整体识别性能的影响，每次增加相同的步长，见图 4。从曲线的趋势变化可以看出，识别准确率随着阈值的增加在提升，然而在超过某个临界值时又开始下降。在 UCF Sports 数据集上 $T_s=50$ 和 YouTube 数据集上 $T_s=50$ 分别取得了最好的结果，故取 $T_s=50$ 作为显著性阈值。

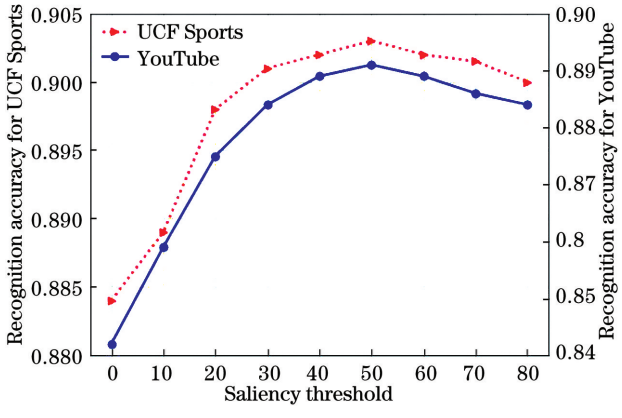


图 4 显著性区域检测参数的估计

Fig. 4 Estimation of saliency detection parameters

### 3.3 实验结果

由于传统稠密轨迹算法只考虑了特征的几何特

性，未考虑显著性，所以并不能很好地区分动作区域和背景运动区域。如图 5(a)～(b)所示，给出本文方法与稠密轨迹方法在两个数据集上(骑马和跳水动作)的可视化比较。通过显著图的指示，将轨迹的追踪确定在动作相关区域。如图 5(b)所示，依据背景运动的相似性和对应矩阵具有的低秩特性，得知背景运动存在于一个低秩的子空间中，通过优化(3)式求解出表示背景运动的低秩矩阵，所以视频中的相机运动和背景目标的运动(指泳池中水的流动)可被大部分滤除。为了验证本文显著性稠密轨迹(S-Traj)的有效性，在两个数据集上比较了显著性稠密轨迹和稠密轨迹(DT)对于每个动作类的识别精度。如图 6(a)所示，UCF Sports 数据集上显著性稠密轨迹在所有动作类中均有较好的识别效果，其中有 7 个类(golf swinging，kicking，riding horse，running，skateboarding，swing bench 和 swing side)的识别率高于稠密轨迹。在图 6(b)中的结果表明，显著性稠密轨迹在 YouTube 数据集上同样取得了较好的识别性能，其中有 9 个类(shooting，biking，diving，horse back riding，soccer juggling，swinging，tennis swinging，volleyball spiking 和 walking with a dog)的识别率高于稠密轨迹。

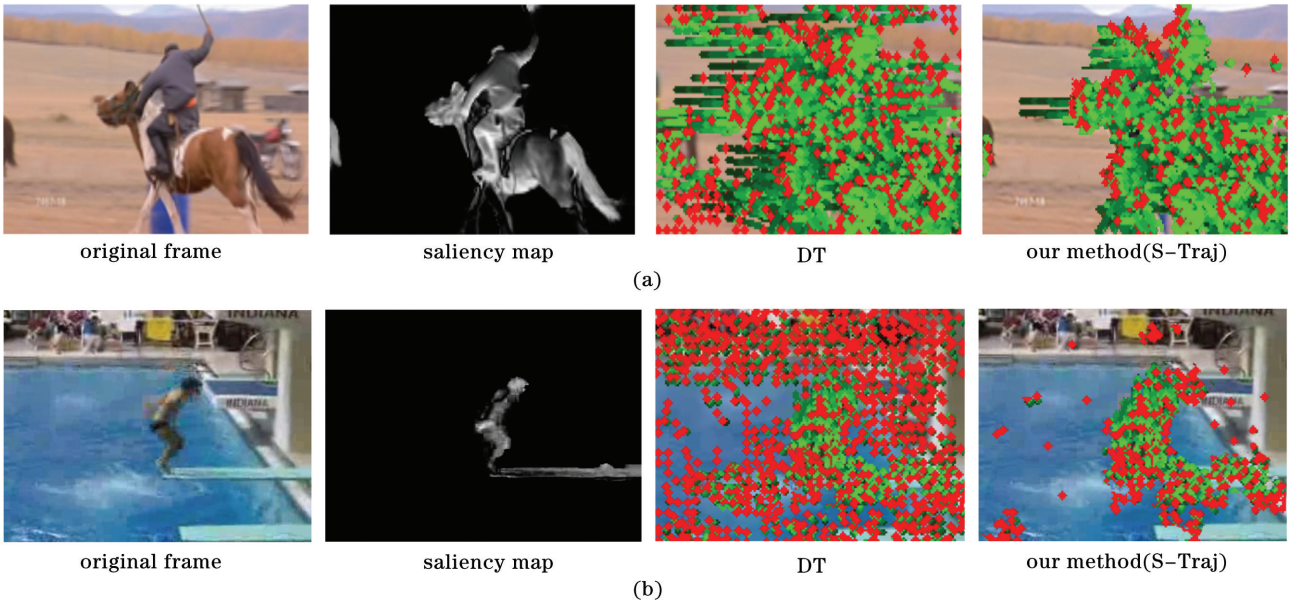


图 5 本文方法与稠密轨迹方法的可视化比较。(a)UCF Sports 数据集；(b)YouTube 数据集

Fig. 5 Comparison of the DT and our method. (a) UCF Sports; (b) YouTube

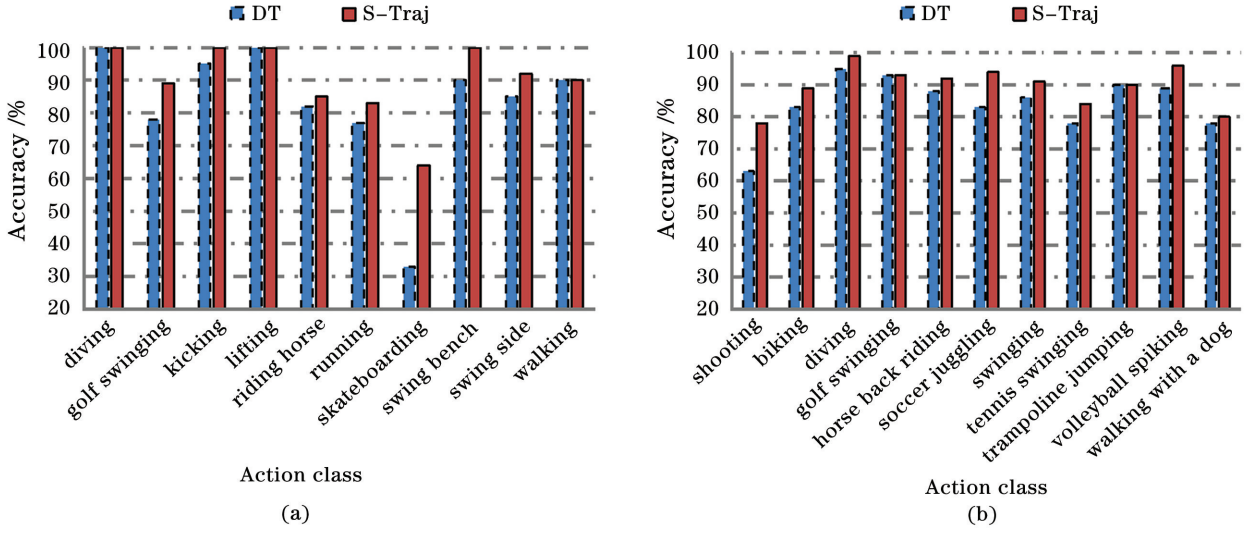


图 6 本文方法与稠密轨迹方法对每个动作类识别率的比较。(a)UCF Sports 数据集；(b)YouTube 数据集

Fig. 6 Accuracy comparison of each class by DT and our method. (a) UCF Sports; (b) YouTube

表 2 将本文方法与传统的稠密轨迹平均识别率进行比较。本文方法在 UCF Sports 数据集中的平均识别准确率高于稠密轨迹方法 2.1 个百分点，在 YouTube 数据集中要高于稠密轨迹方法 5.5 个百分点。

表 2 本文方法与传统稠密轨迹方法平均识别准确率的比较

Table 2 Comparison of mean accuracy by DT and our method unit: %

| Datasets | DT | S-Traj |
|---|---|---|
| UCF Sports | 88.2 | 90.3 |
| YouTube | 84.1 | 89.6 |

表 3 对比了本文方法与近几年相关的动作识别方法在两个数据集上的平均识别准确率。在 UCF Sports 数据集中，本文所提方法的识别准确率要比文献[6]中使用投影变换矩阵来消除相机运动的方法高 1.2 个百分点，比文献[13]中联合表观显著性和运动显著性去除冗余轨迹的方法高 0.22 个百分点，比文献[14]中使用稀疏表示全局时空特征的方法高 3 个百分点，比文献[21]中利用局部运动和群稀疏性的动作识别方法高 0.6 个百分点。在 YouTube 数据集中，本文方法的识别精度要比文献[22]中利用超稀疏编码向量对运动信息、表观信息和位置信息联合建模的方法高 1.6 个百分点，比文献[23]中基于运动边界采样和 3D 共生描述子的方法高 2 个百分点，比文献[24]中通过迭代训练来筛选具有区分性的超体素进行视频表示的方法高 0.1 个百分点。然而，在 UCF Sports 数据集上本文方法的识别精度低于文献[15]所提出的利用运动显著性值提纯轨迹和优化词袋模型的方法；同样在 YouTube 数据集上略低于文献[25]中融合视觉显著性选择判别性区域表示动作的方法，这也从另一个角度阐明了利用显著性来优化轨迹是一种切实可行且有效的方式。最后，深度网络与手动提取特征的方法相结合[26-27]，形成优势互补，在目前的大规模行为识别数据集上性能表现出色，这给我们未来探索更具有表现力和区分度的特征提供了一种可行的思路。

表 3 本文方法与当前先进算法实验结果的对比

Table 3 Results comparison of our method and the state-of-the-art method unit: %

| UCF Sports | | YouTube | |
|---|---|---|---|
| Method | Mean accuracy | Method | Mean accuracy |
| Wang *et al*[6] | 89.10 | Wang *et al*[6] | 85.40 |
| Yi *et al*[13] | 90.08 | Yang *et al*[22] | 88.00 |
| Somasundaram *et al*[14] | 87.30 | Peng *et al*[23] | 87.60 |
| Li *et al*[15] | 93.40 | Guo *et al*[24] | 89.50 |
| Cho *et al*[21] | 89.70 | Duan *et al*[25] | 90.00 |
| Our method | **90.30** | Our method | **89.60** |

## 4 结 论

为了有效地处理传统稠密轨迹表示冗余性的问题，提出了一种新的稠密轨迹采样策略(S-Traj)，基于低秩分解的思想，利用时空信息将视频分解为存在背景运动的低秩矩阵和与动作相关的稀疏误差矩阵，使用得到的稀疏矩阵计算视频显著图并根据显著图指示的区域提取轨迹。实验结果表明，对视频中背景运动带来的干扰进行了较好的处理，节省了内存的开销。在两个公开的数据集上本文的方法取得了较好的识别性能，与传统的稠密轨迹算法相比，识别准确率取得了一定程度的提升。本文方法虽然对传统的稠密轨迹做出了更有效的表示，但是由于对显著性区域的分析，增加了算法的时间复杂度，未来的研究将主要降低算法复杂度、提高算法普适性。

## 参 考 文 献

[1] Luo H L, Wang C J, Lu F. Survey of video behavior recognition[J]. Journal on Communications, 2018, 39(6): 169-180.
罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 169-180.

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

[3] Laptev I. On space-time interest points[J]. International Journal of Computer Vision, 2005, 64(2/3): 107-123.

[4] Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]// 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, October 15-16, 2005, Beijing, China. New York: IEEE Press, 2005: 65-72.

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

[6] Wang H, Schmid C. Action recognition with improved trajectories[C]//2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 3551-3558.

[7] Liu X, Zhao G Y, Yao J W, et al. Background subtraction based on low-rank and structured sparse decomposition[J]. IEEE Transactions on Image Processing, 2015, 24(8): 2502-2514.

[8] Souly N, Shah M. Visual saliency detection using group lasso regularization in videos of natural scenes[J]. International Journal of Computer Vision, 2016, 117(1): 93-110.

[9] Li Y D, Xu X P. Video saliency detection method based on spatiotemporal features of superpixels[J]. Acta Optica Sinica, 2019, 39(1): 0110001.
李艳荻, 徐熙平. 基于超像素时空特征的视频显著性检测方法[J]. 光学学报, 2019, 39(1): 0110001.

[10] Duan L J, Xi T, Cui S, et al. A spatiotemporal weighted dissimilarity-based method for video saliency detection[J]. Signal Processing: Image Communication, 2015, 38: 45-56.

[11] Li Q W, Zhou Y Q, Ma Y P, et al. Salient object detection method based on binocular vision[J]. Acta Optica Sinica, 2018, 38(3): 0315002.
李庆武, 周亚琴, 马云鹏, 等. 基于双目视觉的显著性目标检测方法[J]. 光学学报, 2018, 38(3): 0315002.

[12] Wang L, Zhao D B. Recognizing actions using salient features[C]//2011 IEEE 13th International Workshop on Multimedia Signal Processing, October 17-19, 2011, Hangzhou, China. New York: IEEE Press, 2011: 1-6.

[13] Yi Y, Lin Y K. Human action recognition with salient trajectories[J]. Signal Processing, 2013, 93(11): 2932-2941.

[14] Somasundaram G, Cherian A, Morellas V, et al. Action recognition using global spatio-temporal features derived from sparse representations[J]. Computer Vision and Image Understanding, 2014, 123: 1-13.

[15] Li Q, Cheng H, Zhou Y, et al. Human action recognition using improved salient dense trajectories[J]. Computational Intelligence and Neuroscience, 2016, 2016: 6750459.

[16] Yi Y, Zheng Z X, Lin M Q. Realistic action recognition with salient foreground trajectories[J]. Expert Systems With Applications, 2017, 75: 44-55.

[17] Wang X F, Qi C. Saliency-based dense trajectories for action recognition using low-rank matrix decomposition[J]. Journal of Visual Communication and Image Representation, 2016, 41: 361-374.

[18] Wang X F, Qi C, Lin F. Combined trajectories for action recognition based on saliency detection and motion boundary[J]. Signal Processing: Image Communication, 2017, 57: 91-102.

[19] Rodriguez M D, Ahmed J, Shah M. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE Press, 2008: 1-8.

[20] Liu J G, Luo J B, Shah M. Recognizing realistic actions from videos "in the wild"[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 1996-2003.

[21] Cho J, Lee M, Chang H J, et al. Robust action recognition using local motion and group sparsity[J]. Pattern Recognition, 2014, 47(5): 1813-1825.

[22] Yang X D, Tian Y L. Action recognition using super sparse coding vector with spatio-temporal awareness[C]// Fleet D, Pajdla T, Schiele B, et al. Computer Vision-ECCV 2014. Cham: Springer, 2014: 727-741.

[23] Peng X J, Qiao Y, Peng Q. Motion boundary based sampling and 3D co-occurrence descriptors for action recognition[J]. Image and Vision Computing, 2014, 32(9): 616-628.

[24] Guo Y N, Ma W, Duan L J, et al. Human action recognition based on discriminative supervoxels[C]// 2016 International Joint Conference on Neural Networks (IJCNN), July 24-29, 2016, Vancouver,

BC, Canada. New York: IEEE Press, 2016: 3863-3869.

[25] Duan L J, Guo Y N, Qiao Y H, et al. Human action recognition based on extracted discriminative regions [J]. Journal of Beijing University of Technology, 2017, 43(10): 1480-1487.
段立娟, 郭亚楠, 乔元华, 等. 基于判别性区域提取的视频人体动作识别方法[J]. 北京工业大学学报, 2017, 43(10): 1480-1487.

[26] Wang L M, Qiao Y, Tang X O. Action recognition with trajectory-pooled deep-convolutional descriptors [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 4305-4314.

[27] Li Q H, Li A H, Wang T, et al. Double-stream convolutional networks with sequential optical flow image for action recognition[J]. Acta Optica Sinica, 2018, 38(6): 0615002.
李庆辉, 李艾华, 王涛, 等. 结合有序光流图和双流卷积网络的行为识别[J]. 光学学报, 2018, 38(6): 0615002.