

6-2-2020

## Action Recognition by Improved Dense Trajectories

Peizhen Xu

*School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China;*

Zhibin Yu

*School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China;*

Weidong Jin

*School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China;*

Haiying Jiang

*School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China;*

Follow this and additional works at: <https://dc-china-simulation.researchcommons.org/journal>



Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, Numerical Analysis and Scientific Computing Commons, Operations Research, Systems Engineering and Industrial Engineering Commons, and the Systems Science Commons

---

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

---

## Action Recognition by Improved Dense Trajectories

### Abstract

**Abstract:** In recent years, with strong robustness to fast irregular motion, the method of human motion representation based on dense trajectories has been used more and more in the field of behavior recognition. However, the relative motion of the background caused by the motion of the camera has a great influence on the extraction of the trajectories. *In order to estimate the camera motion, the speed up robust feature (SURF) descriptor was used to match the feature points of each frame. Since the human motion and the camera motion were not same, human detection was added to remove inconsistent matches.* Finally, multi instance learning (MIL) was used to classify and recognize the behavior. Experiment results demonstrate the effectiveness of the approach on the UT-interaction dataset.

### Keywords

dense trajectories, the estimate of camera motion, human detection, multi instance learning

### Recommended Citation

Xu Peizhen, Yu Zhibin, Jin Weidong, Jiang Haiying. Action Recognition by Improved Dense Trajectories[J]. Journal of System Simulation, 2017, 29(9): 2053-2058.

# 基于提高的稠密轨迹人体行为识别

许培振, 余志斌, 金炜东, 蒋海英

(西南交通大学电气工程学院, 四川 成都 610031)

**摘要:** 由于稠密轨迹对快速不规则运动的鲁棒性较强, 近年来基于稠密轨迹的人体运动表征方法越来越多地运用于行为识别等领域。但是由于相机运动造成背景相对运动, 对轨迹的提取产生了很大的影响。加入了相机运动估计, 为了估计相机运动, 用快速鲁棒特征描述符匹配每一帧的特征点。由于人的运动和相机运动不一样, 加上了人体检测去除不一致的匹配。通过多示例学习对交互行为进行分类识别。并在 UT-Interaction 数据集上进行了测试, 证明了该方法的有效性。

**关键词:** 稠密轨迹; 相机运动估计; 人体检测; 多示例学习

中图分类号: TP391 文献标识码: A 文章编号: 1004-731X (2017) 09-2053-06

DOI: 10.16182/j.issn1004731x.joss.201709023

## Action Recognition by Improved Dense Trajectories

Xu Peizhen, Yu Zhibin, Jin Weidong, Jiang Haiying

(School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China)

**Abstract:** In recent years, with strong robustness to fast irregular motion, the method of human motion representation based on dense trajectories has been used more and more in the field of behavior recognition. However, the relative motion of the background caused by the motion of the camera has a great influence on the extraction of the trajectories. In order to estimate the camera motion, the speed up robust feature (SURF) descriptor was used to match the feature points of each frame. Since the human motion and the camera motion were not same, human detection was added to remove inconsistent matches. Finally, multi instance learning (MIL) was used to classify and recognize the behavior. Experiment results demonstrate the effectiveness of the approach on the UT-interaction dataset.

**Keywords:** dense trajectories; the estimate of camera motion; human detection; multi instance learning

## 引言

近几年, 随着监控系统的迅速发展, 需要处理的视频量也成指数增长, 其中大多数记录人类社会活动。为了能达到智能监控, 因此, 能准确地识别视频中人体的动作具有重要意义<sup>[1]</sup>。

人体动作识别在过去今年已成为热门研究领

域, 在公共数据集上也得到了很好的识别结果。但在真实情况下, 由于存在背景复杂、摄像机运动、遮挡和光线变化等影响因素, 对动作识别产生了很大影响。因此, 急需解决由于摄像机运动而产生的背景运动, 而现有的工作还鲜有人研究这方面的问题。

一些研究者已经提出很多方法, 并且都得到了很好的识别结果。如: 时空兴趣点与词袋模型相结合, 虽然可以对视频图像的关键点进行统计, 但是却忽略了兴趣点的局部区域和空间信息之间的联系。还有在建立视觉码本的时候, 需要 K-means



收稿日期: 2017-05-20 修回日期: 2017-07-17;  
基金项目: 国家自然科学基金(61461051), 国家科技支撑计划(2015BAG14B01-05);  
作者简介: 许培振(1991-), 男, 河南商丘, 硕士, 研究方向为计算机视觉; 余志斌(1976-), 男, 湖南, 博士, 副教授, 研究方向为信号处理。

<http://www.china-simulation.com>

• 2053 •

聚类形成词袋模型, 这样会损失很多重要的细节信息。轨迹方法已在文献[2]中提出, 但是它容易受到相机和背景运动的影响, 会把由于相机运动产生的轨迹也提取出来, 这部分轨迹会产生一定的干扰, 影响动作的识别<sup>[3-7]</sup>。

## 1 提高稠密轨迹

在本节中, 首先描述了相机运动估计方法的主要步骤, 以及如何使用它来改善密集的轨迹。然后, 讨论如何消除潜在的不符合基于人类获得的单应性矩阵鲁棒估计。

为了估计全背景运动, 假设连续两帧的单应性相关<sup>[8]</sup>。这种假设在大多数情况下是成立的, 因为两帧之间的全局运动通常是很小的。不包括独立的移动物体, 例如人和车辆。

为了估计单应性矩阵, 第一步是找到两帧之间的对应关系。结合两种方法, 以产生足够的、互补的候选匹配。然后提取 SURF 特征<sup>[9]</sup>并用最邻近原则匹配他们, 之所以选择 SURF 特征是因为他们对运动模糊具有一定的鲁棒性, 就像文献[10]评价的那样。

为了能够提供两帧之间的稠密匹配, 从光流场采样运动向量。在这里, 使用一个基于多项式展开高效的光流算法<sup>[11]</sup>, 利用良好的特征跟踪标准选择运动矢量的显著特征点<sup>[12]</sup>, 例如, 设定自相关矩阵最小特征值阈值。这两种方法是互补的, SURF 特征注重团状结构, 而文献[12]注重棱角和边界。

然后用 RANSAC 估计单应性矩阵<sup>[13]</sup>, 用于删除相机运动。对于密集的轨迹, 从光流场取消相机运动主要有两个优点。首先, 运动描述符可以直接获得, 像文献[14]说的那样, HOF 描述符显著降低相机运动的运动存在。其次, 可以删除摄像机运动所产生的轨迹, 这可以通过设定轨迹位移矢量的阈值获得。如果位移矢量过小, 这个轨迹就可以认为是由于相机运动产生的, 那么就删除它, 这样就消除由于人产生的不一致匹配。

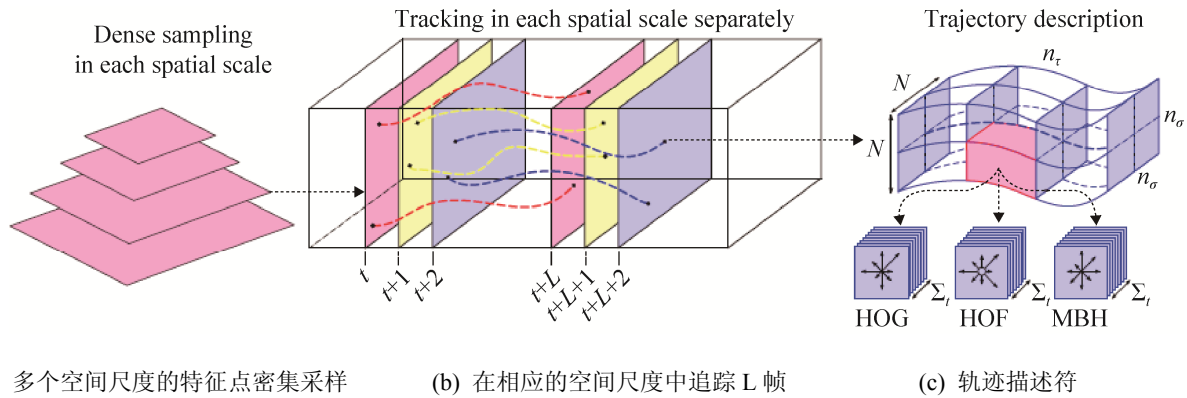
在动作数据集中, 动作往往是人执行的。所以人在每一帧中都会出现, 这对相机运动估计是一个干扰, 因为人类的运动一般不符合它。所以我们使用一个人体检测器, 以消除在人类区域的匹配。一般情况下, 在行动数据集上的人体检测是相当困难的, 因为人正在执行行动时往往有很大的姿势变化。

这里, 我们用一个最先进的人体监测<sup>[15]</sup>的方法, 它在一般的动作数据库中能很好地检测出人体区域。该检测器可以结合人体不同部位的特征。当估计单应性矩阵<sup>[16]</sup>时, 去除用人体区域的特征点匹配, 以提高单应性矩阵的准确率。人体检测器并不是一直表现的很好, 由于姿势和视角的变化有可能丢失检测, 为了补偿丢失检测, 我们追踪所有检测到的边界盒子, 对视频的每一帧进行向前和向后追踪, 每个边界盒子最多追踪 15 帧, 如果和其它边界盒子有 50% 的重叠部分就停止。以后, 在计算单应性矩阵之前, 总是先用人体检测消除潜在的不一致匹配。

## 2 稠密轨迹的特征描述符

稠密轨迹的提取是在多尺度空间下进行的, 文献[6]提出, 在相应的尺度空间下对像素点跟踪  $L$  帧, 像素点的领域大小为  $n \times n$ , 这样就形成了大小为  $n \times n \times L$  的是视频快, 为了获取周围信息, 轨迹邻域被分为  $n_{\sigma} \times n_{\sigma} \times L_{\tau}$  个时空网格, 然后再提取每个网格的特征, 如图 1 所示。在视频的每一帧图像中, 采用多尺度空间进行采样, 由于图像的每个像素之间的联系, 所以需要像素领域进行联合编码, 这样才能获得更多的细节信息。因此本文选取时间共现描述符对轨迹进行描述。

To-moki Watanabe 等人<sup>[17]</sup>提出了空间共现方向梯度直方图(Co HOG), 最早应用于作行人检测。Peng 等人<sup>[18]</sup>扩展了 Co HOF 与 Co MBH, 把它用来描述行为特征。



(a) 多个空间尺度的特征点密集采样 (b) 在相应的空间尺度中追踪 L 帧 (c) 轨迹描述符

图 1 稠密轨迹描述符的说明

Fig. 1 Illustration of our dense trajectory description

为了能从上下文中获得清晰的运动信息和外观变化, 需要 3 个时间共现情境描述符, 分别为 T-Co HOG, T-Co HOF 与 T-Co MBH。T-Co HOG 是由成对的方向梯度构成的, 在一定的时间偏移量中, 通过中值滤波光流跟踪两帧像素形成共生矩阵, 它用来描述形状信息。T-Co HOF 是由光流方向时间对形成的, 它用来描述运动特征。T-Co MBH 是由水平和垂直方向的成对的光流梯度构成的, 它用来描述运动边界的局部变化。

由图 2 所示, (a)是成对表示的 T-Co HOG 和 T-Co MBH, (b)是成对表示的 T-Co HOF。并且可以看出至少需要 3 帧图像才能获得时间共现描述符的基本单元。随着时间变化, T-Co HOG 能够描述外部形状的改变。类似地, T-Co HOF 能够描述运动方向的改变, T-Co MBH 能够描述光流梯度方向的改变。

轨迹的形状需要对局部运动模式进行编码, 这个轨迹特征向量用  $S'$  表示, 有形状序列:

$$s = (\Delta P_t, \dots, \Delta P_{t+L-1}) \quad (1)$$

位移矢量:

$$\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t) \quad (2)$$

$$S' = \frac{(\Delta P_t, \dots, \Delta P_{t+L-1})}{\sum_{j=1}^{t+L-1} \|\Delta P_j\|} \quad (3)$$

其中:  $L$  表示轨迹的长度;  $s$  表示一段序列的形状;  $\Delta P_t$  表示一段时间后的位移矢量<sup>[19]</sup>。

综上所述, 一段视频可以由很多轨迹经过聚类

后进行表示, 而时空情景描述符又可以很好的描述这些轨迹。

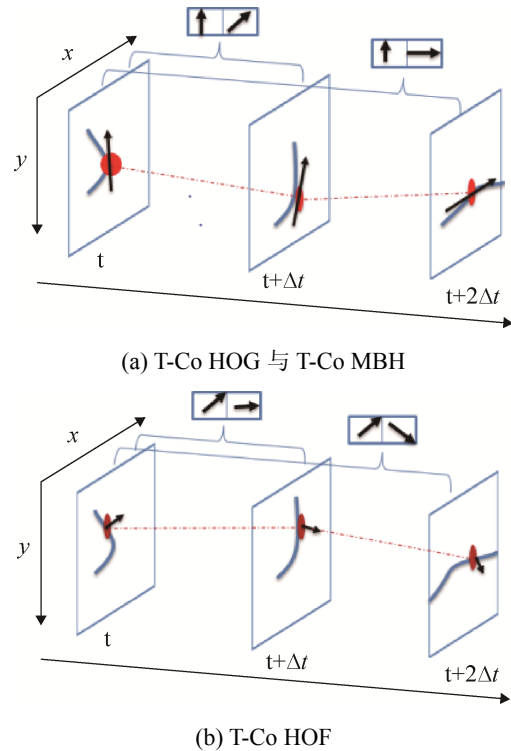


图 2 时间共现描述符

Fig. 2 Temporal co-occurrence descriptors

### 3 多示例学习

一般的轨迹方法用于动作识别, 都是把运动描述符与词袋模型相结合<sup>[18]</sup>, 例如, 在 Wang 等<sup>[2]</sup>的方法中, 他们首先计算每条轨迹的 HOF、MBH 和 HOG, 然后再用 K-means 对训练集中所有描述符

进行聚类,产生视觉码本。每个动作的特征根据欧式距离归类到离其最近的码本中,形成一个标准化的特征向量。最后使用标准 SVM 进行分类。该模型虽然取得了不错的效果,但存在一定的缺陷,一是当数据的背景复杂时,词袋模型效果不好,二是使用 K-means 聚类,鲁棒性较低,易受数据的数量和特征的影响,并且  $K$  值需要靠经验获得,导致码本尺寸难以确定。

本文使用基于神经网络的 MIL 算法 (BP-MIP)<sup>[20]</sup>,假设有  $N$  包  $\{B_1, B_2, \dots, B_N\}$ ,第  $i$  个包含有  $M_i$  示例  $\{B_{i1}, B_{i2}, \dots, B_{iM_i}\}$ ,每个示例是个  $p$  维的特征向量,例如,第  $i$  个包的第  $j$  个示例表示为  $[B_{ij1}, B_{ij2}, \dots, B_{ijp}]^T$ ,假设有一个前馈神经网络,有  $P$  个输入单元,一个输出单元,和一个隐藏层。

由于多示例学习是预测未知包的标签,定义一个多示例误差函数代替标准 BP 使用的误差函数,

$$E_i = \begin{cases} \min_{1 \leq j \leq M_j} E_{ij} & \text{if } B_i = + \\ \max_{1 \leq j \leq M_j} E_{ij} & \text{if } B_i = - \end{cases} \quad (4)$$

式中:“ $B_i = +$ ”(“ $B_i = -$ ”)表示  $B_i$  是正包;  $E_{ij}$  是示例  $B_{ij}$  的误差:

$$E_{ij} = \frac{1}{2}(o_{ij} - d_i)^2 \quad (5)$$

式中:  $o_{ij}$  是第  $i$  个训练包的第  $j$  个示例在 BP 神经网络上的输出;  $d_i$  为训练集中第  $i$  包的期望输出。在多示例学习分类器中,含有包和示例,包由示例组成。包有标记,而示例没有标记。多示例学习的规则是,有一个示例被分类器标签为+,则该包为正包,只有所有的示例被分类器标记为-,该包才为负包。

## 4 实验数据和环境

在本文试验中,我们采用文献[21]中提供的 UT-Interaction 数据库,部分动作如下图所示,这个数据库包含 6 类交互动作,分别为:握手、拥抱、踢人、指人、拳击和推人。这个数据集又分为两个子数据集: SET1 和 SET2,其中 SET1 的背景是停车场,背景几乎是完全静止的,且相机几乎没有运

动;而 SET2 是在有风的草地上拍摄的,背景是动态的,在拍摄过程中相机是运动的。两个场景的背景和拍摄条件完全不同。在本实验中,用的是 8GRAM、2.8 GHz Intel Core i7 处理器华硕笔记本电脑, MATLAB 平台。

本文实验参数的选取参考文献[6],即轨迹长度  $L=15$ ,时空网格大小为  $n_{\tau}=3, n_{\delta}=2$ ,像素领域  $N=32$ ,采样步长为  $W=5$ 。

本文在 UT-Interaction 数据集上做了测试,实验结果如图 3 所示,在 SET1 数据集上除了“拥抱”和“拳击”识别率基本达到 100%,其余各个动作的识别率相对均衡,都达到了 90%左右。在 SET2 上的识别率和 SET1 相差无几,由于 SET2 数据库的拍摄包含很多相机运动和背景运动,这也说明了加入了相机估计后确实克服了由于相机的运动对识别率造成的影响。

Handshaking	0.94	0.00	0.00	0.00	0.06	0.00
Hugging	0.00	1.00	0.00	0.00	0.00	0.00
Kicking	0.00	0.00	0.89	0.00	0.01	0.10
Pointing	0.10	0.00	0.00	0.90	0.00	0.00
Punching	0.00	0.00	0.00	0.00	1.00	0.00
Pushing	0.00	0.00	0.00	0.00	0.10	0.90

(a) SET1 的混淆矩阵

Handshaking	0.94	0.00	0.00	0.00	0.06	0.00
Hugging	0.00	0.95	0.05	0.00	0.00	0.00
Kicking	0.00	0.00	0.95	0.00	0.00	0.05
Pointing	0.05	0.05	0.00	0.90	0.00	0.00
Punching	0.00	0.00	0.00	0.00	0.97	0.03
Pushing	0.10	0.00	0.00	0.00	0.00	0.90

(b) SET2 的混淆矩阵

图 3 SET1 和 SET2 的混淆矩阵

Fig. 3 Confusion matrix of SET1 and SET2

本文还对比了前人的方法分别在 SET1 和 SET2 子数据上的识别率。其中 Ryoo M S<sup>[22]</sup>等人应用时空关系匹配法,就是在两个视频中提取特征结构相似性,最后用 SVM 进行分类; Alonso P P<sup>[23]</sup>采用头部方向和它们周围的局部时空区域为特征描述符,采用把局部和全局描述符相结合的结构化的输出支持向量机作为学习和推理模型; Yang 等人<sup>[24]</sup>则把关键点速度随时间的变化为特征,并用此特征生成混合序列模型。实验结果表明,他们的方法在 SET1 和 SET2 数据库上的识别率是有差别的,在 SET2 上的识别率比 SET1 的识别率要低 3% 左右,这也说明了他们的方法在应对相机运动上效果较差,但本文的方法在两个子数据集上的识别率基本没有变化,这也说明了此方法的有效性。

表 1 UT-Interaction 子数据集上准确率的对比  
Tab. 1 Comparison of accuracy rates on subdata sets of UT-Interaction /%

方法	SET1	SET2	平均
Ryoo M S	93	90	92
Alonso P P	84	82	83
Yang	96.7	93.3	95
本文方法	93.8	93.5	93.6

## 5 结论

本文针对现有方法对有相机运动的动作数据库识别率不高的问题,研究了具有相机运动估计的稠密轨迹算法,具体过程是用 RANSAC 估计单应性矩阵,用现在效果最好的人体检测器,去除相机运动估计不一致的匹配,这可以帮助我们删除由于相机运动导致背景运动产生的轨迹。此外,本文采用时空情景描述符 Co MBH、Co HOG 和 Co HOF,相比于 MBH、HOF 和 HOG 描述符能更好的表述特征。本文采用多示例学习分类器进行分类,相比于 SVM 分类器运算速度更快,精度更高。

## 参考文献:

[1] Zhu J, Zhao Y, Tang J. Automatic Recognition of Radar Signals Based on Time-frequency Image Character [J].

- Defence Science Journal (S0011-748X), 2013, 63(3): 1-6.
- [2] Wang H, Klaser A, Schmid C, et al. Action Recognition by Dense Trajectories [C]// Proceeding of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2011: 3169-3176.
- [3] Aggarwal J K. Recognition of Human Activities [M]// Combinatorial Image Analysis. Berlin, Heidelberg, Germany: Springer, 2011: 1-4.
- [4] Laptev I, Lindeberg T. On Space-Time Interest Points [J]. International Journal of Computer Vision (S0920-5691), 2005, 64(2/3): 107-123.
- [5] 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. USA: IEEE Computer Society, 2005: 65-72.
- [6] Messing R, Pal C, Kautz H. Activity Recognition Using the Velocity Histories of Tracked Keypoints [C]// Proceedings IEEE International Conference on Computer Vision. USA: IEEE, 2009: 104-111.
- [7] Matikainen P, Hebert M, Sukthankar. Trajectons: Action Recognition Through the Motion Analysis of Tracked Features [C]// Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. USA: IEEE, 2009: 514-521.
- [8] Wang H, Klaser A, Schmid C, et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition [J]. International Journal of Computer Vision (S0920-5691), 2013, 103(1): 60-78.
- [9] Szeliski R. Image Alignment and Stitching: a Tutorial [J]. Foundations & Trends in Computer Graphics & Vision (S1572-2740), 2004, 2(1): 101-104.
- [10] Bay H, Tuytelaars T, Gool L V. SURF: Speeded Up Robust Features [J]. Computer Vision & Image Understanding (S1077-3142), 2006, 110(3): 404-417.
- [11] Gaidon A, Harchaoui Z, Schmid C. Recognizing Activities with Cluster-trees of Track lets [C]// British Machine Vision Conference. United Kingdom: BNVA Press, 2012: 1-6.
- [12] Farneback G. Two-frame motion estimation based on polynomial expansion [C]// Scandinavian Conference on Image Analysis. Germany: Springer-Verlag, 2003: 363-370.
- [13] Shi J, Tomasi C. Good Features to Track [M]. USA: Cornell University, 1993.
- [14] Fischler M A, Bolles R C. Rand on Sample Consensus: a Paradigm for Model Fitting with Applications to Image

