

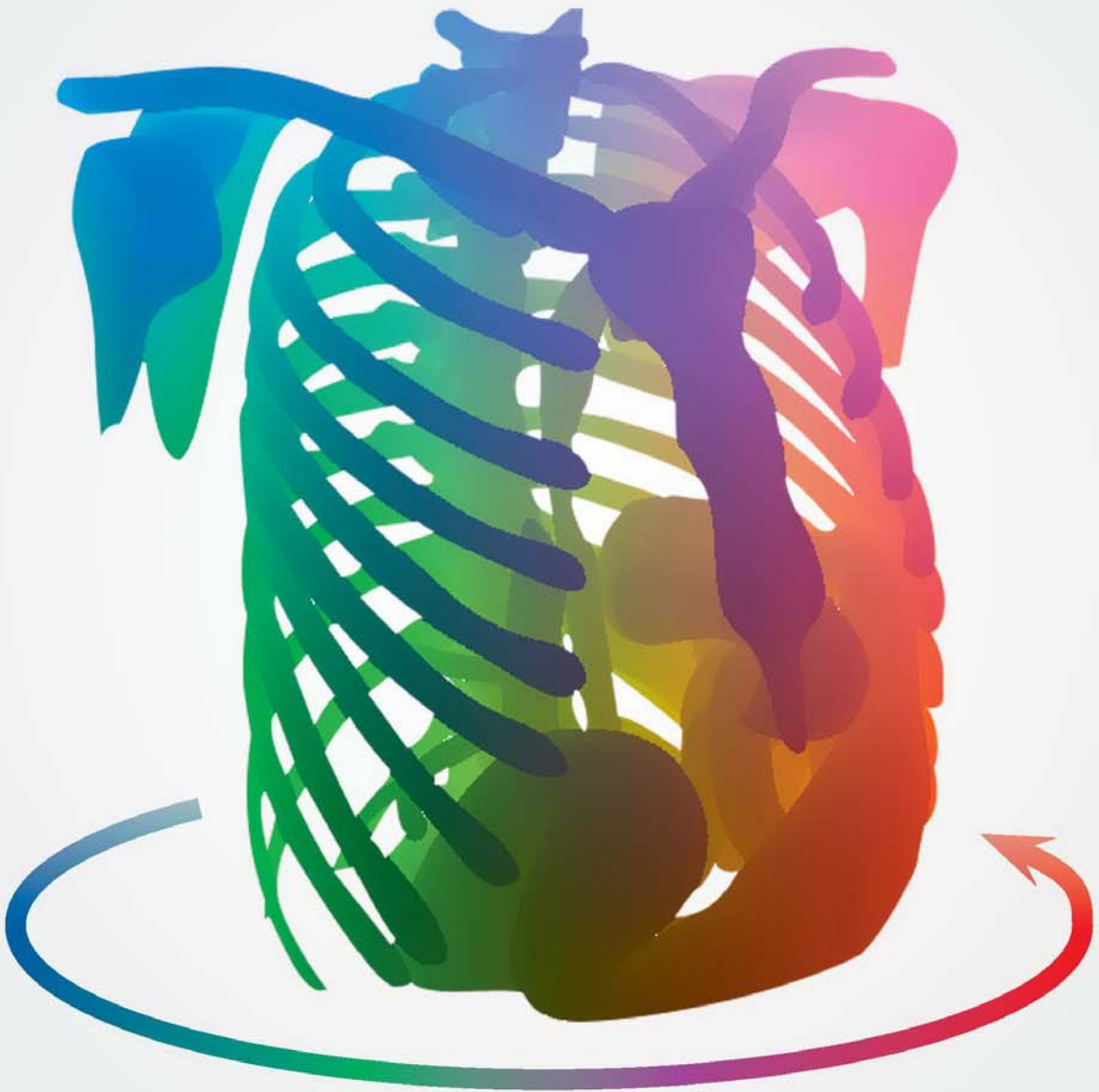


主办: 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2015
02
VOL.20

ISSN1006-8961
CN11-3758/TB



等值面光线跟踪 P193



第20卷第2期（总第226期）

2015年2月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院

主办单位 中国科学院遥感与数字地球研究所

中国图象图形学学会

北京应用物理与计算数学研究所

主 编 李小文

编辑出版《中国图象图形学报》编辑出版委员会

邮政信箱 北京9718信箱

邮 编 100101

电子信箱 jig@irsa.ac.cn

电 话 010-64807995

网 址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号

总 发 行 北京报刊发行局

订 购 全国各地邮局

海外发行 中国国际图书贸易集团有限公司

（邮政信箱：北京399信箱 邮编：100048）

印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian

Monthly, Started in 1996

Superintended by Chinese Academy of Sciences

Sponsored by Institute of Remote Sensing and Digital Earth, CAS

China Society of Image and Graphics

Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen

Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics

P.O.Box 9718, Beijing, P.R.China

Zip code 100101

E-mail jig@radi.ac.cn

Telephone 010-64807995

Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestic All Local Post Offices in China

Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China))

Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB

ISSN 1006-8961

CODEN ZTTXFZ

国外发行代号 M1406

国内邮发代号 82-831

国内定价 50.00元



多尺度活动网格在云场景仿真中的应用(第0202页)



自然环境视频中基于显著鲁棒轨迹的行为识别(第0245页)



虚拟手抓持力觉生成算法真实性的评价(第0280页)

图像处理和编码

引入神经网络中间神经元的快速小波图像压缩

张海涛, 张永霖 0159

拟正态分布扩散的图像平滑

周先春, 汪美玲, 周林锋 0169

可选特征的快速分形图像编码

袁宗文, 鲁业频, 杨汉生 0177

图像分析和识别

融合多尺度码本的全局编码图像分类

董振宇, 赵杰煜, 祝军 0183

计算机图形学

结合K-D树和Shell的快速动态等值面光线跟踪法

罗月童, 石放放, 张伟, 朱会国 0193

多尺度活动网格在云场景仿真中的应用

范晓磊, 张立民, 张建廷, 徐涛 0202

第十届和谐人机环境联合会议专栏

低资源语言的无监督语音关键词检测技术综述

杨鹏, 谢磊, 张艳宁 0211

快速离散Curvelet变换域的图像融合

杨勇, 董松, 黄淑英 0219

高效视频编码中变换跳过模式的快速选择

王宁, 张永飞, 樊锐 0229

主动学习的多标签图像在线分类

徐美香, 孙福明, 李豪杰 0237

自然环境视频中基于显著鲁棒轨迹的行为识别

易云, 王瀚漓 0245

特征变换和数据集分块的行人比对

沈洋, 林巍晓, 殷惠清 0254

移动手持环境下触摸式目标选择技术对比

辛义忠, 李岩, 袁伟强, 郑谦 0264

虚拟环境的用户意图捕获

程成, 赵东坡, 卢保安 0271

虚拟手抓持力觉生成算法真实性的评价

杨文珍, 许艳, 颜传武, 吴新丽, 张昊, 孟闯 0280

交互式乐器演奏的六自由度力觉渲染方法

童号, 史有姣, 王党校, 张玉茹 0288

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



Multi-scale moving grids and its application in clouds scene simulation(P0202)



Action recognition from unconstrained videos via salient and robust trajectory (P0245)



Validation the Authenticity of Virtual fingers Grasping Haptic Algorithm(P0280)

Image Processing and Coding

- Fast wavelet image compression introducing neural network of relay neuron
Zhang Haitao, Zhang Yonglin 0159
- Image smoothing algorithm based on matching normal distribution diffusion
Zhou Xianchun, Wang Meiling, Zhou Linfeng 0169
- Optional features fast fractal image coding algorithm
Yuan Zongwen, Lu Yepin, Yang Hansheng 0177

Image Analysis and Recognition

- Image classification based on global coding combined with multi-scale codebook
Dong Zhenyu, Zhao Jieyu, Zhu Jun 0183

Computer Graphics

- Fast isosurface ray tracing method by combining K-D tree and Shell
Luo Yuetong, Shi Fangfang, Zhang Wei, Zhu Huiguo 0193
- Multi-scale moving grids and its application in clouds scene simulation
Fan Xiaolei, Zhang Limin, Zhang Jianting, Xu Tao 0202

Column of HHME'2014

- Survey on unsupervised spoken term detection for low-resource languages
Yang Peng, Xie Lei, Zhang Yanning 0211
- Image fusion based on fast discrete Curvelet transform
Yang Yong, Tong Song, Huang Shuying 0219
- Fast transform skip mode decision for high efficiency video coding
Wang Ning, Zhang Yongfei, Fan Rui 0229
- Online multi-label image classification with active learning
Xu Meixiang, Sun Fuming, Li haojie 0237
- Action recognition from unconstrained videos via salient and robust trajectory
Yi Yun, Wang Hanli 0245
- Pedestrian re-identification based on feature transformation and dataset classification
Shen Yang, Lin Weiyao, Yin Huiqing 0254
- Comparison of target selection techniques in mobile handheld environment
Xin Yizhong, Li Yan, Yuan Weiqiang, Zheng Qian 0264
- Intent understanding for virtual environments
Cheng Cheng, Zhao Dongpo, Lu Baoan 0271
- Validation the authenticity of virtual fingers grasping haptic algorithm
Yang Wenzhen, Xu Yan, Yan Chuanwu, Wu Xinli, Zhang Hao, Meng Chuang 0280
- Interactive simulation of music instrument playing by a six degree-of-freedom haptic rendering approach
Tong Hao, Shi Youjiao, Wang Dangxiao, Zhang Yuru 0288

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2015)02-0245-09

论文引用格式: Yi Y, Wang H L. Action recognition from unconstrained videos via salient and robust trajectory[J]. Journal of Image and Graphics, 2015, 20(2): 0245-0253. [易云, 王瀚漓. 自然环境视频中基于显著鲁棒轨迹的行为识别[J]. 中国图象图形学报, 2015, 20(2): 0245-0253.][DOI:10.11834/jig.20150211]

自然环境视频中基于显著鲁棒轨迹的行为识别

易云^{1,2}, 王瀚漓¹

1. 同济大学计算机科学与技术系, 上海 201804; 2. 赣南师范学院数学与计算机科学系, 赣州 341000

摘要: **目的** 人类行为识别是计算机视觉领域的一个重要研究课题。由于背景复杂、摄像机抖动等原因,在自然环境视频中识别人类行为存在困难。针对上述问题,提出一种基于显著鲁棒轨迹的人类行为识别算法。**方法** 该算法使用稠密光流技术在多尺度空间中跟踪显著特征点,并使用梯度直方图(HOG)、光流直方图(HOF)和运动边界直方图(MBH)特征描述显著轨迹。为了有效消除摄像机运动带来的影响,使用基于自适应背景分割的摄像机运动估计技术增强显著轨迹的鲁棒性。然后,对于每一类特征分别使用 Fisher Vector 模型将一个视频表示为一个 Fisher 向量,并使用线性支持向量机对视频进行分类。**结果** 在 4 个公开数据集上,显著轨迹算法比 Dense 轨迹算法的实验结果平均高 1%。增加摄像机运动消除技术后,显著鲁棒轨迹算法比显著轨迹算法的实验结果平均高 2%。在 4 个数据集(即 Hollywood2、YouTube、Olympic Sports 和 UCF50)上,显著鲁棒轨迹算法的实验结果分别是 65.8%、91.6%、93.6% 和 92.1%,比目前最好的实验结果分别高 1.5%、2.6%、2.5% 和 0.9%。**结论** 实验结果表明,该算法能够有效地识别自然环境视频中的人类行为,并且具有较低的时间复杂度。

关键词: 行为识别;显著轨迹;摄像机运动消除;Fisher vector

Action recognition from unconstrained videos via salient and robust trajectory

Yi Yun^{1,2}, Wang Hanli¹

1. Department of Computer Science and Technology, Tongji University, Shanghai 201804, China;

2. Department of Mathematics and Computer Science, Gannan Normal University, Ganzhou 341000, China

Abstract: **Background** In the past few years, we have witnessed a great success of social networks and multimedia technologies, leading to the generation of vast amount of Internet videos. To organize these videos and to provide value-added services to users, human activities from videos should be automatically recognized. A number of research studies have focused on this challenging topic. **Objective** Human action recognition is a significant research topic in computer vision. The recognition of human actions from unconstrained videos is difficult because of complex background and camera motion. A robust and salient trajectory-based approach is proposed to address such problem. **Method** Dense optical flow is utilized to track the scale invariant feature transform keypoints at multiple spatial scales. The histogram of oriented gradient, histogram of optical flow, and

收稿日期:2014-07-25;修回日期:2014-09-23

基金项目:国家自然科学基金项目(61102059,61472281);中央高校基本科研业务费专项资金(0800219158,0800219270);上海市曙光计划(12SG23)

第一作者简介:易云(1983—),男,现为同济大学计算机科学与技术专业博士研究生,主要研究领域为计算机视觉和机器学习。E-mail: 13yiyun@tongji.edu.cn

通信作者:王瀚漓,E-mail: hanliwang@tongji.edu.cn

Supported by: National Natural Science Foundation of China (61102059, 61472281); Fundamental Research Funds for the Central Universities (0800219158, 0800219270)

motion boundary histogram are employed to depict the trajectory efficiently. To eliminate the influence of camera motions, a camera motion estimation approach based on adaptive background segmentation is utilized to improve the robustness of trajectory. The Fisher vector model is utilized to compute one Fisher vector over a complete video for each descriptor separately, and the linear support vector machine is employed for classification. **Result** The average improvement of salient trajectory algorithm over dense trajectory algorithm is 1% on four challenging datasets. After utilizing the camera motion elimination approach, the average experimental result over salient trajectory is improved by 2%. The state-of-the-art results on four datasets (*i. e.*, Hollywood2, YouTube, Olympic Sports and UCF50), the proposed algorithm obtains 65.8%, 91.6%, 93.6%, 92.1%, and the state-of-the-art results have been improved by 1.5%, 2.6%, 2.5%, 0.9% respectively. **Conclusion** Experimental results on four challenging datasets demonstrate that the proposed algorithm can effectively recognize human actions from unconstrained videos in a more computationally efficient manner compared with a number of state-of-the-art approaches.

Key words: action recognition; salient trajectory; camera motion elimination; Fisher vector

0 引言

随着多媒体技术与计算机网络的广泛应用,网络上出现大量视频数据。为了能够有效地管理这些视频文件,并且给用户提供更好的体验服务,需要一种技术能够准确、高效地识别自然环境(即无约束环境)视频中的人类行为。

自然环境下拍摄的视频与约束环境下拍摄的视频相比,具有背景复杂、摄像机运动、物体遮挡等特点。目前,已经有一些研究工作尝试解决自然环境中的人类行为识别问题^[1-6],但是还存在以下问题需要解决:

1) 视频中人类的行为可能被摄像机运动混淆,这将干扰识别人类行为。

2) 算法的特征计算过程需要消耗大量时间,具有较高的计算复杂度。

3) 在某些数据集(如:Hollywood2^[1])上的人类行为识别准确率不高。

为了解决上述问题,提出基于显著鲁棒轨迹的人类行为识别方法。该方法通过稠密光流^[7]技术在多尺度空间上跟踪显著点,并使用摄像机运动估计技术减少摄像机运动对行为识别造成的影响。为了能够有效地描述轨迹特征,采用基于轨迹的梯度直方图特征^[8](HOG)、光流直方图特征^[9](HOF)和运动边界直方图特征^[10](MBH)分别描述人类行为的静态和动态特征。实验结果显示,本文提出的方法在4个公开数据集上的识别率高于目前所公布的其他方法,且具有较低的计算复杂度。

目前存在很多从视频中提取运动特征的技术,其中基于轨迹的技术是研究的热点之一。根据形成

轨迹的方式,可以将基于轨迹的技术分为3类:1)使用特征点匹配技术形成轨迹;2)使用稀疏光流^[11]技术形成轨迹;3)使用稠密光流^[7]技术形成轨迹。

Matikainen 等人^[12]提出使用稀疏光流 KLT tracker^[11]技术形成轨迹。Sun 等人^[13]提出使用 SIFT^[14](scale invariant feature transform)特征点匹配技术形成轨迹。Wang 等人^[6]提出使用稠密光流技术形成轨迹。

有效地使用轨迹技术跟踪人类行为存在两个关键点。首先是选取合适的轨迹起始点;其次是选择合适的跟踪轨迹点的方法。

由于 SIFT 特征点^[14]具有高稳定性和消除边界响应的特点,本文选择 SIFT 特征点作为轨迹的起始点。由于稠密光流^[7]能够很好地描述每一个点的运动,本文在多尺度空间中使用稠密光流技术跟踪每一个 SIFT 特征点。文献[13]也使用了 SIFT 特征点,但是跟踪过程使用的是 SIFT 特征点匹配技术。本文首次提出在多尺度空间中,使用稠密光流技术跟踪 SIFT 特征点,实验证明该方法能够较好地识别人类行为。

文献[15]使用行人检测技术进行摄像机运动估计,并且取得了较好的实验结果。由于行人检测本身是一个比较难的问题^[15],并且具有较高的计算复杂度,所以本文提出基于自适应背景分割^[16]的摄像机运动估计方法。实验证明,该方法能够有效地消除摄像机运动对行为识别造成的影响。

1 显著鲁棒轨迹

提出一种基于显著鲁棒轨迹的人类行为识别算法。该算法框架如图1所示。

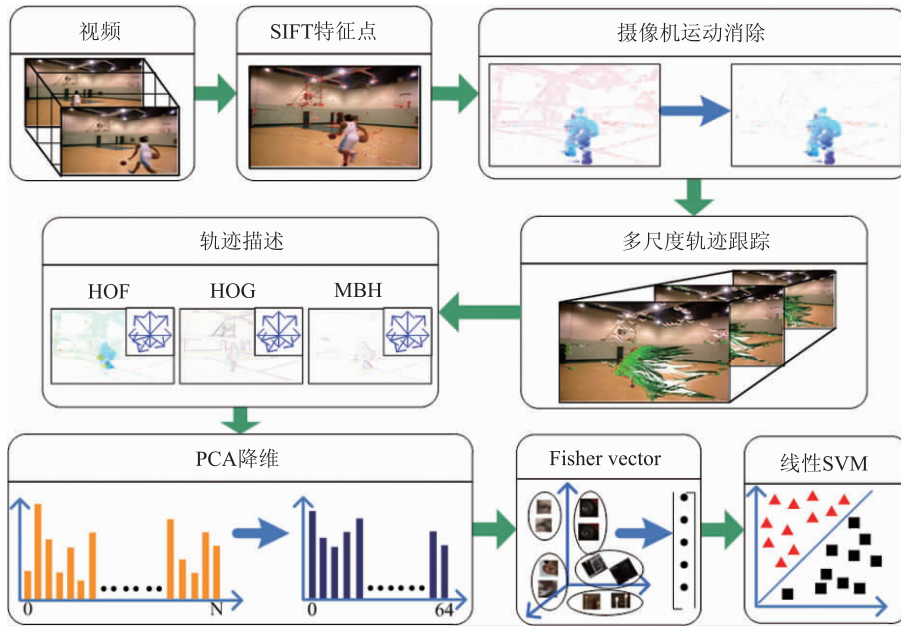


图1 算法框架图

Fig. 1 Overview of the algorithm

1.1 显著轨迹

轨迹指视频序列中点的有序集合。设存在某一个起始点 (x_0, y_0) , 则以该点为起始点的轨迹定义为 $T = \{(x_0, y_0), (x_1, y_1), \dots, (x_{L-1}, y_{L-1})\}$, 其中 L 是轨迹的长度。

在实验中, 以 $\sqrt{2}$ 为缩放因子, 将坐标点投影到尺度空间 s , 即

$$[x_s, y_s]^T = \left[\frac{x}{\sqrt{2^{s-1}}}, \frac{y}{\sqrt{2^{s-1}}} \right]^T \quad (1)$$

$$[x_{i+1}, y_{i+1}]^T = [x_i, y_i]^T + (\mathbf{F} \cdot \mathbf{M}) \mid_{[x_i, y_i]^T} \quad (2)$$

式中, $\mathbf{F}$ 表示当前帧的稠密光流矩阵, $\mathbf{M}$ 表示中值滤波核。式(2)是计算某轨迹中最后一个点 (x_i, y_i) 在当前帧中的位置。

选择 SIFT 特征点作为轨迹的起始点, 显著轨迹的具体形成过程如算法 1 所示。实验结果表明, 选择 SIFT 特征点作为轨迹的起始点能够有效地提升算法的识别率。

算法 1 显著轨迹形成算法 TRACK

输入: 一个视频的帧序列 I , 总帧数 N , 轨迹长度 L 。

输出: 轨迹的特征。

TRACK(I, N, L)

BEGIN

定义 2 维数组 $\mathbf{V}$ 用来存储轨迹;

计算第 1 帧图像的 SIFT 特征点 $\mathbf{P}$;

以 $\sqrt{2}$ 为缩放因子, 计算最大的尺度空间数 MS ;

FOR($s = 1; s \leq MS; s++$)

FOR EACH $p(x, y) \in \mathbf{P}$

根据式(1)计算 p 在尺度 s 的坐标 $p_s(x_s, y_s)$;

以起始点 $p_s(x_s, y_s)$ 初始化轨迹 K ;

初始化 K 的长度为 1;

将 K 加入尺度 s 的轨迹数组 $\mathbf{V}[s]$ 中;

END FOR

END FOR

FOR($t = 2; t \leq N; t++$)

计算当前帧图像的 SIFT 特征点 $\mathbf{P}$;

FOR($s = 1; s \leq MS; s++$)

计算当前帧在尺度 s 上的稠密光流矩阵 $\mathbf{F}$;

对矩阵 $\mathbf{F}$ 进行中值滤波;

FOR EACH $K \in \mathbf{V}[s]$

计算 K 的长度 i ;

IF $i = L$ THEN

输出轨迹的特征;

从尺度 s 的轨迹数组 $\mathbf{V}[s]$ 中删除该轨迹;

ELSE

根据式(2)计算 K 中最后一个点 (x_i, y_i) 在当前帧的位置 (x_{i+1}, y_{i+1}) ;

将 (x_{i+1}, y_{i+1}) 加入轨迹 K ;

计算当前帧的轨迹特征;

轨迹 K 的长度加 1;

END IF

END FOR

FOR EACH $p(x, y) \in \mathbf{P}$

根据式(1)计算 p 在尺度 s 的坐标 $p_s(x_s, y_s)$;
 计算 $V[s]$ 中所有轨迹的最后一个点的集合 Q ;
 IF Q 中所有点与 $p_s(x_s, y_s)$ 的距离都大于 5, THEN
 以起始点 $p_s(x_s, y_s)$ 初始化轨迹 K ;
 初始化 K 的长度为 1;
 将 K 加入尺度 s 的轨迹数组 $V[s]$ 中;
 END IF
 END FOR
 END FOR

END FOR
 END

1.2 摄像机运动消除

在拍摄视频过程中,摄像机运动将使相对静止的背景物体变成运动的物体。为了消除摄像机运动对行为识别产生的影响,提出一种基于背景分割的摄像机运动估计算法。该算法的核心思想是找到背景中前后两帧的匹配点,然后通过 Random sample consensus^[17] 估计摄像机的运动变换矩阵。

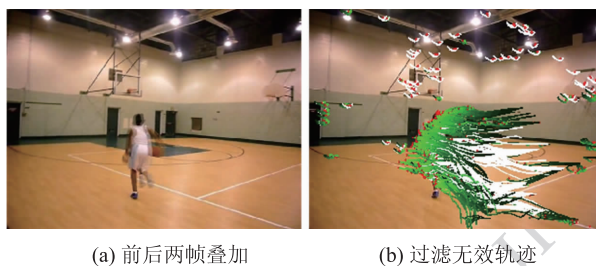
首先,使用文献[16]中的背景分割算法计算每一帧的背景区域。基于像素的自适应分隔算法是一种非参数的背景分割算法,文献[16]通过实验证明该算法优于目前其他类似算法。该算法的主要思想是使用具有循环反馈的控制器去控制置信值阈值和学习参数。

其次,使用 SIFT 特征点匹配方法,在背景区域内计算前后两帧之间匹配的特征点数组 P_s 。为了增加匹配的特征点数,与文献[15]类似,本文通过 Shi-Tomasi 角点^[18] 和稠密光流矩阵 F 计算前后两帧的角点匹配数组 P_c 。在算法实现中,该过程重复利用算法 1 中计算的 SIFT 特征点和光流矩阵,使得减少算法的重复计算量。

最后,输入匹配点数组 P_s 和 P_c ,使用 Random sample consensus 方法计算透视变换矩阵 H 。根据 H 重新计算当前帧所有点的位置,从而消除摄像机运动产生的影响。图 2(b) 中白色线条表示被过滤的无效轨迹,说明摄像机运动估计技术可以消除部分背景干扰、增强轨迹的鲁棒性。

1.3 轨迹特征描述

选择 HOG^[8] 描述轨迹的静态特征,选择 HOF^[9] 和 MBH^[10] 描述轨迹的动态特征。HOG 特征表示静态图像的边界统计信息,HOF 特征表示前后帧之间的运动统计信息,MBH 特征表示前后帧之间的运动



(a) 前后两帧叠加

(b) 过滤无效轨迹

图 2 消除摄像机运动效果示意图

Fig. 2 The effect of camera motion elimination

边界统计信息。图 3 是某一帧的上述 3 种特征的形象表示,其中使用颜色表示方向,使用饱和度表示梯度。图 3(c) 中背景区域的光流是由于摄像机运动产生。通过对比图 3(c) 与图 3(d) 可以发现,当存在摄像机运动时,HOF 特征会被干扰,而 MBH 特征具有一定的抗干扰能力。



(a) 原始图片

(b) HOG特征

(c) HOF特征

(d) MBH特征

图 3 某一帧的特征示意图

Fig. 3 Schematic diagram of features

本文设置轨迹的参数与文献[6]一致,如轨迹长度为 15 帧。HOG、HOF 和 MBH 特征的维度分别是 96、108 和 192。与文献[15]类似,对于每一类特征,本文先使用 L1 归一化,然后再对每一维单独计算平方根。

2 行为分类

2.1 Fisher vector 模型

Fisher vector 模型^[19] 是 Bag-of-visual-words (BOW)模型的扩展。文献[19]通过实验证明在解决图像分类问题时 Fisher vector 模型比 BOW 模型

效果更好。文献[5,15]将该模型应用在视频分类中,并通过实验证明在解决行为识别问题时 Fisher vector 模型也比 BOW 模型效果更好。

为了减少计算复杂度,与文献[5]类似,本文使用主成分分析方法(PCA)将每一个特征描述向量降维成 $D = 64$ 维。在所有实验中,均设置固定的高斯数 $K = 500$ 。对 HOG、HOF 和 MBH 特征分别计算视觉词汇表,然后分别使用一个长度为 $2DK$ 的 Fisher 向量表示一个视频。最后,使用平方根归一化和 L2 归一化方法对每一个 Fisher 向量进行归一化操作。

2.2 线性支持向量机

文献[19]中提出,线性支持向量机(linear SVM)^[20]更适合 Fisher vector 模型,可以用较小的计算代价获得较好的实验结果。

在实验中,使用前融合方法线性组合 HOG、HOF 和 MBH 特征;使用 one-against-rest 方法解决多分类问题;使用对正、负例样本进行加权的方法实现样本平衡,从而解决训练集的正负例样本不平衡问题。

3 实验分析

3.1 实验数据集

在4个公开的人类行为数据集(即 Hollywood2^[1]、YouTube^[21]、Olympic Sports^[22] 和 UCF50^[23])上验证算法的行为识别效果。这些数据集中的视频都是在自然环境下拍摄,主要来源于好莱坞电影或者 YouTube 视频分享网站,总共有大约 1 万个视频和 89 类人类行为。图4是这4个数据集中某些人类行为的图像示例。



图4 数据集中部分行为的图像示例

Fig. 4 Sample frames from the datasets

Hollywood2^[1]数据集包含12类行为。在实验中,使用该数据集的 clean 子集,其中包含823个训练视频和884个测试视频。与文献[1]一致,实验输出每一类的平均精度均值(mean average precision)。

YouTube^[21]数据集包含11类行为,总共有

1 168 个视频。与文献[21]一致,实验使用留一交叉验证方法(leave-one-out cross-validation),输出每一类的平均准确率(average accuracy)。

Olympic Sports^[22]数据集包含16类行为,总共有783个视频。与文献[22]一致,实验输出每一类的平均精度均值。

UCF50^[23]数据集包含 50 类行为,共有 6 618 个视频。与文献[23]一致,实验使用留一组交叉验证方法(leave one group out cross-validation),输出每一类的平均准确率。

3.2 与 Dense 轨迹的比较

为了验证显著轨迹算法的识别效果,选择 Dense 轨迹算法作为参考标准,分别在 4 个数据集上进行比较。实验过程中,显著轨迹与 Dense 轨迹都使用相同的轨迹参数。表 1 是显著轨迹与 Dense 轨迹算法进行比较的实验结果。表 1 中 HOG、HOF 和 MBH 表示分别只使用对应特征时的实验结果,组合表示将 3 类特征进行线性连接后的实验结果。

表 1 比较显著轨迹与 Dense 轨迹

方法 特征		平均识别率/%			
		Hollywood2	YouTube	Olympic	UCF50
Dense 轨迹	HOG	48.6	85.6	79.8	83.9
	HOF	53.3	80.0	78.9	80.9
	MBH	59.2	86.6	85.1	87.3
	组合	62.0	89.5	89.5	89.8
显著 轨迹	HOG	46.4	83.4	80.4	83.5
	HOF	53.8	81.5	86.5	81.6
	MBH	60.6	88.3	88.0	88.4
	组合	62.9	90.0	91.3	90.8

对于 HOG 静态特征,使用 Dense 轨迹算法的识别率高于显著轨迹。这是由于对于静态特征,选择的点越多,具有越强大的表示能力^[6]。Dense 轨迹算法使用稠密取点,使得轨迹数明显多于显著轨迹算法。所以对于 HOG 特征,在 4 个数据集上,Dense 轨迹算法的平均识别率比显著轨迹高 1%。

对于运动特征(即 HOF 和 MBH),使用显著轨迹算法的识别率高于 Dense 轨迹。在 4 个数据集上,对于 HOF 特征,显著轨迹算法比 Dense 轨迹算法平均识别率高 2.6%;对于 MBH 特征,显著轨迹算法比 Dense 轨迹算法平均识别率高 1.8%。这是由于运动特征的识别效果与轨迹起始点选取的位置有关。计算 HOF 和 MBH 特征时,以轨迹点为中心

计算运动特征。如果计算中心在运动物体的边缘,则基于直方图的特征无法很好地表示运动信息;如果计算中心在运动物体的中心,则基于直方图的特征能够较好地表示运动信息。SIFT 特征点能够消除边缘响应^[14],所以使用运动特征时显著轨迹算法的识别率高于 Dense 轨迹。

在 4 个数据集上,线性组合 HOG、HOF 和 MBH 特征后,显著轨迹算法比 Dense 轨迹算法的识别率平均高 1%。这说明显著轨迹算法比 Dense 轨迹算法具有更强的人类行为表示能力。

3.3 摄像机运动消除前后的比较

为了评估摄像机运动消除技术对显著轨迹算法带来的影响,本文选择显著轨迹算法作为参考标准。在 4 个数据集上,分别使用 HOG、HOF、MBH 及其组合特征计算显著轨迹算法和显著鲁棒轨迹算法(即:增加了摄像机运动消除技术的显著轨迹算法)的实验结果。在实验过程中,显著轨迹算法与显著鲁棒轨迹算法使用相同的参数。表 2 是增加摄像机运动消除技术前后实验结果的比较。

表 2 增加摄像机运动消除技术的影响

方法 特征		平均识别率/%			
		Hollywood2	YouTube	Olympic	UCF50
显著 轨迹	HOG	46.4	83.4	80.4	83.5
	HOF	53.8	81.5	86.5	81.6
	MBH	60.6	88.3	88.0	88.4
	组合	62.9	90.0	91.3	90.8
显著 鲁棒 轨迹	HOG	48.1	85.1	82.9	83.9
	HOF	58.1	86.3	90.3	85.3
	MBH	64.1	89.5	91.4	89.2
	组合	65.8	91.6	93.6	92.1

对于 HOG 静态特征,由于使用摄像机运动消除技术后,能够剔除一些由于摄像机运动产生的轨迹。所以,显著鲁棒轨迹算法比显著轨迹算法的实验结果平均提升 1.5%。

使用摄像机运动消除技术后,对动态特征(即 HOF 和 MBH 特征)的影响非常显著。相对 MBH 特

征, 使用该技术后, 对 HOF 特征的识别率提升得更多。在 4 个数据集上, 对于 HOF 特征, 使用该技术后的识别率平均提升 4.2%; 对于 MBH 特征, 识别率平均提升 2.2%。这是由于 MBH 特征表示的是运动边界特征, 能够消除部分摄像机运动对行为识别带来的影响, 所以使用该技术后 MBH 特征的识别率提升低于 HOF 特征。

在 4 个数据集上, 使用摄像机运动消除技术后线性组合 3 类特征的识别率平均提升 2%。实验证明, 本文提出的摄像机运动估计算法能够有效地提升人类行为识别的准确率。

3.4 时间复杂度比较

为了评估显著轨迹算法的时间复杂度, 本文选择 Hollywood2 数据集中 700 个训练视频作为实验测试数据, 使用 Dense 轨迹和显著轨迹算法分别计算这 700 个视频的 HOG、HOF 和 MBH 特征。实验分别统计这两种算法处理视频所花费的平均时间。实验在一台主频是 3.4 GHz 并且内存是 8 G 的电脑上进行。在实验过程中, 设置这两种算法的轨迹参数相同, 并且算法实现都没有使用并行处理技术, 只使用一个 CPU 核心。

图 5 是 Dense 轨迹与显著轨迹算法的时间复杂度比较示意图。其中, $\Delta T_{N,abs}$ 表示提取 N 个视频的特征时 Dense 轨迹比显著轨迹算法多花费的绝对时间, $\Delta T_{N,\%}$ 表示两者的相对时间差。即

$$\Delta T_{N,abs} = \frac{T_N^d - T_N^s}{N}$$
$$\Delta T_{N,\%} = \frac{T_N^d - T_N^s}{T_N^d} \times 100\%$$

式中, N 是视频数, T_N^d 表示使用 Dense 轨迹算法处理 N 个视频花费的时间, T_N^s 表示使用显著轨迹算法

处理 N 个视频花费的时间。

如图 5 所示, 显著轨迹算法在本实验中所花费的时间明显少于 Dense 轨迹。这是因为使用显著轨迹算法提取特征时, 需要计算、存储的轨迹特征数目远少于 Dense 轨迹算法。文献[6]中通过实验表明轨迹特征的计算和存储过程都比较花时间。

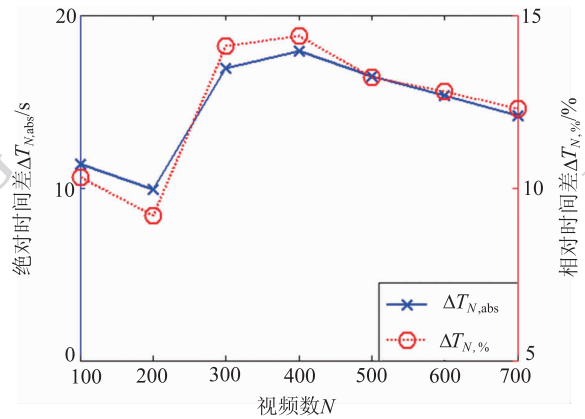


图 5 时间复杂度比较示意图

Fig. 5 Comparison of computational complexity

3.5 与最新成果的比较

表 3 是本文提出的显著鲁棒轨迹算法与最近两年相关算法的实验结果的比较。其中, 文献[2]使用一种局部的轨迹特征描述视频中的人类行为, 在 YouTube 数据集上得到实验结果 88.0%。文献[3]使用 Divergence-curl-shear 特征和 VLAD 模型描述人类行为, 在 Hollywood2 和 Olympic Sports 数据集上分别得到实验结果 62.5% 和 83.2%。文献[4]使用一种 STAOG 模型描述行为的潜在结构, 在 YouTube 数据集上得到实验结果 88.9%。文献[6]使用 Dense 轨迹描述人类行为, 在 UCF50 数据集上得到实验结果 85.6%。

表 3 与最新实验结果的比较

Table 3 Comparison with state-of-the-art approaches

Hollywood2		YouTube		Olympic Sports		UCF50	
方法	平均精度均值	方法	平均准确率	方法	平均精度均值	方法	平均准确率
JJB ^[3]	62.5	BDP ^[2]	88.0	JJB ^[3]	83.2	WKS ^[6]	85.6
OVS ^[5]	63.3	LLC ^[4]	88.9	OVS ^[5]	84.6	OVS ^[5]	90.0
WS ^[15]	64.3	OVS ^[5]	89.0	WS ^[15]	91.1	WS ^[15]	91.2
本文	65.8	本文	91.6	本文	93.6	本文	92.1

文献[5]使用基于 Dense 轨迹的特征和 Fisher vector 模型描述人类行为,在 YouTube 数据集上获得目前最好的实验结果 89.0%。文献[15]使用基于行人检测技术的摄像机运动估计方法改进 Dense 轨迹,在 Hollywood2、Olympic Sports 和 UCF50 数据集上获得目前最好的实验结果,分别是 64.3%、91.1% 和 91.2%。这些实验数据,从一个侧面说明基于轨迹的方法是当前人类行为识别领域中一种非常有效的方法。

在 Hollywood2、YouTube、Olympic Sports 和 UCF50 数据集上,本文提出的显著鲁棒轨迹算法的实验结果分别是 65.8%、91.6%、93.6% 和 92.1%,比目前最好的实验结果分别高 1.5%、2.6%、2.5% 和 0.9%,这是目前在这 4 个数据集上最好的实验结果。

4 结 论

为了快速、有效地识别自然环境视频中的人类行为,本文提出一种新的行为识别算法。该算法使用稠密光流技术在多尺度空间上跟踪 SIFT 特征点。与 Dense 轨迹算法比较,显著轨迹算法具有识别率高、计算复杂度低的特点。为了减少摄像机运动对人类行为识别造成的干扰,本文提出一种新的摄像机运动估计方法,该方法能够有效的提升显著轨迹算法的行为识别准确率。实验结果表明,在 4 个公开数据集上,该算法的识别率高于近期公布的其他算法。

由于该算法需要处理视频的每一帧,所以当视频的分辨率较高时,在现有的硬件基础上难以做到实时计算视频特征。今后的工作主要关注于如何降低算法计算复杂度和如何提升算法识别率。

参考文献 (References)

- [1] Marszalek M, Laptev I, Schmid C. Actions in context[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA; IEEE, 2009: 2929-2936. [DOI: 10.1109/CVPR.2009.5206557]
- [2] Ballas N, Delezoide B, Preteux F. Trajectory signature for action recognition in video[C]//Proceedings of the ACM International Conference on Multimedia. Nara, Japan; ACM, 2012: 1429-1432. [DOI: 10.1145/2393347.2396511]
- [3] Jain M, Jégou H, Bouthemy P. Better exploiting motion for better action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA; IEEE, 2013: 2555-2562. [DOI: 10.1109/CVPR.2013.330]
- [4] Liang X, Lin L, Cao L. Learning latent spatio-temporal compositional model for human action recognition[C]//Proceedings of the ACM International Conference on Multimedia. Barcelona, Spain; ACM, 2013: 263-272. [DOI: 10.1145/2502081.2502089]
- [5] Oneata D, Verbeek J, Schmid C. Action and event recognition with Fisher vectors on a compact feature set[C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia; IEEE, 2013: 1817-1824. [DOI: 10.1109/ICCV.2013.228]
- [6] 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. [DOI: 10.1007/s11263-012-0594-8]
- [7] Farnebäck G. Two-frame motion estimation based on polynomial expansion[C]//Proceedings of the Image Analysis. Halmstad, Sweden; Springer, 2003: 363-370. [DOI: 10.1007/3-540-45103-X_50]
- [8] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA; IEEE, 2005, 1: 886-893. [DOI: 10.1109/CVPR.2005.177]
- [9] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA; IEEE, 2008: 1-8. [DOI: 10.1109/CVPR.2008.4587756]
- [10] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance[C]//Proceedings of the European Conference on Computer Vision. Graz, Austria; Springer, 2006: 428-441. [DOI: 10.1007/11744047_33]
- [11] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C]//Proceedings of the International Joint Conference on Artificial Intelligence. Vancouver, Canada; Morgan Kaufmann, 1981, 81: 674-679.
- [12] Matikainen P, Hebert M, Sukthankar R. Trajectons: action recognition through the motion analysis of tracked features[C]//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan; IEEE, 2009: 514-521. [DOI: 10.1109/ICCVW.2009.5457659]
- [13] Sun J, Wu X, Yan S, et al. Hierarchical spatio-temporal context modeling for action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA; IEEE, 2009: 2004-2011. [DOI: 10.1109/CVPR.2009.5206721]

- [14] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110. [DOI: 10.1023/B:VISI.0000029664.99615.94]
- [15] Wang H, Schmid C. Action recognition with improved trajectories [C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013:3551-3558. [DOI: 10.1109/ICCV.2013.441]
- [16] Hofmann M, Tiefenbacher P, Rigoll G. Background segmentation with feedback: the pixel-based adaptive segmenter [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Providence, USA: IEEE, 2012: 38-43. [DOI: 10.1109/CVPRW.2012.6238925]
- [17] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [J]. Communications of the ACM, 1981, 24(6): 381-395. [DOI:10.1145/358669.358692]
- [18] Shi J, Tomasi C. Good features to track [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 1994: 593-600. [DOI: 10.1109/CVPR.1994.323794]
- [19] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE, 2007: 1-8. [DOI: 10.1109/CVPR.2007.383266]
- [20] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: a library for large linear classification [J]. The Journal of Machine Learning Research, 2008, 9: 1871-1874.
- [21] Liu J, Luo J, Shah M. Recognizing realistic actions from videos "in the wild" [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009: 1996-2003. [DOI: 10.1109/CVPR.2009.5206744]
- [22] Niebles J C, Chen C W, Fei-Fei L. Modeling temporal structure of decomposable motion segments for activity classification [C]//Proceedings of the European Conference on Computer Vision. Crete, Greece: Springer, 2010: 392-405. [DOI: 10.1007/978-3-642-15552-9_29]
- [23] Reddy K K, Shah M. Recognizing 50 human action categories of web videos [J]. Machine Vision and Applications, 2013, 24(5): 971-981. [DOI: 10.1007/s00138-012-0450-4]