中山大学硕士学位论文

**基于显著轨迹和特征加权的视频行为识别研究**

**Action Recognition in Videos Based on**

**Salient Trajectories and Feature Weighting**

学 位 申 请 人：林乙坤

导师姓名及职称：衣杨 副教授

专 业 名 称：计算机软件与理论

研 究 方 向：网络系统及数据的智能信息处理

答辩委员会主席（签名）：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

委员（签名）：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

二零一三年五月

**论文原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日 期： 年 月 日

**学位论文使用授权声明**

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版；有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅；有权将学位论文的内容编入有关数据库进行检索；可以采用复印、缩印或其他方法保存学位论文；可以为建立了馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

保密论文保密期满后，适用本声明。

学位论文作者签名： 导师签名：

日期： 年 月 日 日期： 年 月

**论文题目**：基于显著轨迹和特征加权的视频行为识别研究

**专业：**计算机软件与理论

**硕士生：**林乙坤

**指导老师：**衣杨 副教授

# 摘 要

视频行为识别是计算机视觉领域的热点问题，对该问题的研究有利于视频内容检索、视频安全监控和人机交互等应用的发展。然而，受视频行为歧义性、行为实体差异性以及背景干扰因素的影响，视频行为识别目前仍然是一项复杂且具有挑战性的工作。

本文针对视频中的人体行为识别问题，提出了基于显著轨迹和特征加权的视频行为识别方法。本文的主要工作包括：

首先，提出了基于显著轨迹的特征提取算法，以缓解传统基于轨迹的特征提取方法存在的噪声过多问题。本文分别从静态和动态两个角度定义了轨迹的外观显著性和运动显著性，并将两者融合得到轨迹的组合显著性，通过对几何特征和显著性的分析提取到有效的轨迹特征。其次，利用运动边界核直方图描述显著轨迹，在量化过程使用高斯函数进行平滑操作，以解决传统直方图硬量化的局限性。其三，提出了针对词汇特征的TF-PRF加权方法，以便更好地区分视觉词汇。采用词袋模型将视频转换到视觉词汇的向量空间，针对词汇在正、负类别视频中的分布信息计算TF-PRF权值，并采用基于非线性SVM的分类算法对视频行为分类。最后，在KTH、ADL和UCF三个行为数据集上，对本文提出的方法进行实验验证。结果表明，本文提出的方法在UCF上的结果与目前最优的方法基本一致，而在KTH和ADL上均取得优于其它方法的识别结果。

**关键字**：行为识别，显著轨迹，核直方图，特征加权

**Title:** Action Recognition in Videos Based on Salient Trajectories and Feature Weighting

**Major:** Computer Software and Theory

**Name:** Yikun Lin

**Supervisor:** Associate Professor Yang Yi

# ABSTRACT

The analysis of human actions in video sequences is attracting much attention in the field of computer vision due to a large number of potential applications including content-based video retrieval, video surveillance, and human computer interactions, et al. Since there are difficulties such as ambiguities of actions, variations of the way person behaves, and background clutter, this topic still remains a complex and nontrivial task.

This paper focuses on the research of human action recognition from video sequences and proposes an action recognition framework based on salient trajectories and feature weighting methods. The primary contributions of this paper are summarized as follows:

First, an algorithm for salient-trajectory extraction is proposed to deal with the problem of noise in traditional methods. At this step, appearance and motion saliency are proposed to capture the static and dynamic visual attentions respectively. And to incorporate complementary visual information, both saliency values are combined. Through the analysis of geometry and saliency properties, a compact set of trajectories is obtained. Secondly, motion boundary kernel histogram, which serves as a Gaussian soft-binning variation of the traditional histogram, is applied for the description of salient trajectories. Thirdly, a new feature weighting method named TF-PRF is formulated to distinguish between visual words. A video sequence is transformed into the vector space of visual words through the bag-of-visual-words (BoVW) model, and each visual word is assigned with an appropriate value by the TF-PRF method, which measures the distribution of visual words in positive and negative categories. For classification purpose, an algorithm based on nonlinear SVM is exploited. Lastly, extensive experiments are carried out on the KTH, ADL, and UCF human action datasets to validate the effectiveness of the proposed method. The comparison with state-of-the-art methods shows that the proposed method achieves comparable results on the UCF dataset and superior results on the KTH and ADL datasets.

**Keywords**: Action recognition, salient trajectory, kernel histogram, feature weighting

# 目 录

[第1章 引言 1](#_Toc351900141)

[1.1 研究背景 1](#_Toc351900142)

[1.2 研究现状 2](#_Toc351900143)

[1.3 本文主要工作 3](#_Toc351900144)

[1.4 本文结构 5](#_Toc351900145)

[第2章 视频行为识别方法综述 7](#_Toc351900146)

[2.1 基于全局特征表示的方法 7](#_Toc351900147)

[2.2 基于局部特征表示的方法 10](#_Toc351900148)

[2.3 基于层次特征表示的方法 13](#_Toc351900149)

[2.4 小结 14](#_Toc351900150)

[第3章 基于显著轨迹和特征加权的视频行为识别方法 17](#_Toc351900151)

[3.1 显著轨迹提取 18](#_Toc351900152)

[3.2 核直方图计算 27](#_Toc351900153)

[3.3 特征加权 30](#_Toc351900154)

[3.4 行为分类 35](#_Toc351900155)

[3.5 时间复杂度 36](#_Toc351900156)

[3.6 小结 37](#_Toc351900157)

[第4章 实验结果与分析 39](#_Toc351900158)

[4.1 实验数据 39](#_Toc351900159)

[4.2 实验结果 42](#_Toc351900160)

[4.3 小结 54](#_Toc351900161)

[第5章 总结与展望 57](#_Toc351900162)

[5.1 全文总结 57](#_Toc351900163)

[5.2 工作展望 57](#_Toc351900164)

[参考文献 59](#_Toc351900165)

[攻读硕士学位期间与硕士学位论文相关的研究成果清单 67](#_Toc351900166)

[致 谢 69](#_Toc351900167)

# 第1章 引言

## 1.1 研究背景

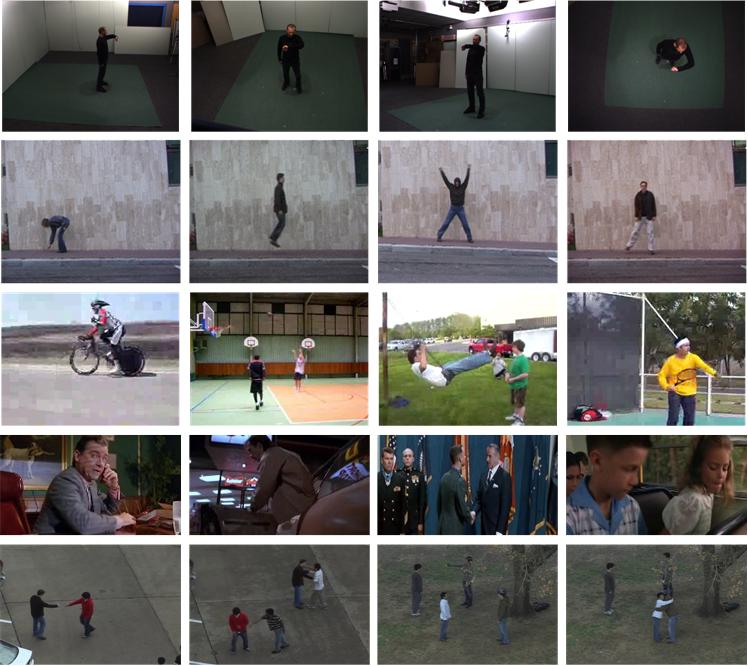
视频行为识别，主要是指运用计算机技术，从已标定类别的视频集中提取特征，对特征进行训练得到分类器，利用分类器来判断未知类别视频的过程。鉴于人体行为对视频内容理解的重要性，现有研究主要集中在人体行为方面。视频行为识别研究综合了多个学科，包括计算机视觉、模式识别、机器学习、数字图像处理等的研究成果。其中，特征表示方式及分类算法在多个领域内具有互相借鉴的意义。由于存在遮挡、拍摄角度/尺寸变化、背景环境干扰和摄像头运动等问题，视频行为识别仍是一项充满挑战的任务。视频行为识别研究的应用前景主要包括：

（1）在视频检索领域的应用。快速发展的互联网技术及存储技术促进了在线视频行业的成长，截至2012年9月，中国在线视频用户规模已经超过4.4亿[[1]](#footnote-1)。因此，快速地在海量视频中检索到用户关注的内容变得尤为重要。现有的搜索引擎将视频检索转化为对人工标注的文本数据的检索，这给视频检索带来了过高的人力成本。要真正实现基于内容的视频检索，需要依靠计算机自动地对视频中的人体行为等内容进行分析。

（2）在视频安全监控中的应用。在商场、医院等公共场所安装摄像头，对犯罪事件及其它突发事件进行监控，有助于营造一个安全和谐的社会环境。现有的视频监控主要依靠人眼进行，需要耗费大量的人力资源。另外，由于人为判断的疏忽，往往达不到理想的监控效果。借助视频行为识别领域的相关技术，可以提高视频监控的可靠性。

（3）在人机交互方面的应用。传统的人机交互主要通过键盘和鼠标等外设进行，这种交互方式和现实世界存在较大差异。另一方面，虚拟现实等技术的发展使得计算机能够更好地理解人类行为。视频行为识别的研究，为未来人类通过摄像头和计算机直接进行交互提供了可能性。

## 1.2 研究现状



**图1-1** 视频行为识别研究的主要数据集（从上到下分别为：监控场景IXMAS[1]、 Weizmann[2], 自然场景Youtube[3]、Hollywood[4]和多实体交互场景UT-Interaction[5]）

视频行为识别的研究，经历了从监控场景到自然场景的发展，以及从单实体场景到多实体交互场景的发展。图1-1展示了各个场景部分数据集的样本，本文采用了监控场景数据集KTH[6]、ADL[7]及自然场景数据集UCF[8]，数据集的样本参照图4-1至图4-3。

前期视频行为识别的研究主要在监控场景进行。监控场景常常对视频的拍摄条件进行约束，如可通过多个摄像头监测人体在立体空间的多角度运动，进而重建行为。另外，监控场景的背景较为固定，能够事先对背景进行建模，以便定位前景运动目标，并持续跟踪目标在各帧的位置。人体形状作为重要的特征广泛应用于该场景。

由于拍摄条件存在限制，监控场景的行为识别方法并不能满足实际应用的需求。因此，自然场景下的视频行为识别研究越来越受到重视。在自然场景下，摄像头可以自由移动，背景环境也可能存在干扰物体，因此该场景下的行为识别研究更具有挑战性。自然场景下难以提取到前景人体形状，主要使用的特征包括兴趣点和运动轨迹等。

单实体行为识别研究取得的进展，推动了多实体交互行为的研究。在交互场景中，需要对交互实体之间的语境（Context）进行建模。目前多实体交互行为的研究还处在起步阶段，因此考虑的场景相对比较简单。

## 1.3 本文主要工作

本文主要研究视频中的单实体行为识别问题。由于不存在对拍摄环境的假设，也不依赖于背景的建模，本文的方法同时适用于监控场景和自然场景。

### 1.3.1 方法动机

如何从视频中提取到一组紧凑且有鉴别性的特征，是视频行为识别研究的首要关注点。现有的特征表示方法主要包括：全局特征表示、局部特征表示和层次特征表示。其中，局部特征是近十年来的研究热点。传统的局部特征表示方法主要通过响应函数来检测时空兴趣点，首先在时空域上的每一点计算特征响应函数，然后取局部最大值对应的点作为检测结果[9,10]。响应函数主要检验各点的几何特征，因此该类方法能提取到位于重要几何位置的特征点，但是却不能保证特征点或者跟踪得到的轨迹与人体行为相关。受到显著性理论相关研究的启发[11, 12]，本文提出了一个基于显著轨迹的特征提取方法，以弥补现有方法的不足。本文的特征提取方法不仅考虑了轨迹的几何意义，而且通过显著值对前景和背景轨迹进行区分。

迄今为止，仅有少数视频行为识别方法将显著性理论融入到特征提取过程中[13-15]。Rapantzikos[13]等人定义了颜色、运动和灰度等特征的相似度，并通过解能量最小化问题来获得各点的显著值。Wang和Zhao[14]在显著图中检测显著区域，为了过滤运动无关区域，使用图模型描述显著区域的分布，并通过图分割算法分割得到特征一致的分量。Vig等人[15]计算区域内的显著性，并使用显著模板来过滤特征描述符。然而，这些方法关注的是特征点或者区域的显著值，没有对轨迹的显著性进行分析。特征点的显著性容易受到噪声的影响，而区域的显著性没有识别个体特征的差异。为了提取到更为可靠的特征，本文对轨迹的显著性进行定义，并将其融入特征提取过程。

在向量空间模型中，一个视频（文本）可以表示成视觉词汇（Visual Word）的特征向量，向量中每个元素分别表示相应词汇的权值。在文本分类问题中，词汇可以对应音节、单词或词组等不同级别的特征，而在视频行为识别问题中，词汇与特征描述符的聚类中心相对应。赋予词汇合适的权值，是特征加权研究的目的。特征加权方法广泛运用于文本分类问题中[16-18]，但是，现有的视频行为识别工作并没有对特征加权方法进行研究，而是直接以词汇在视频中的出现频率作为词汇的权值。本文将特征加权方法引入视频行为识别问题，并通过实验比较多个特征加权方法的识别结果。

视频行为识别属于有监督学习的范畴，训练数据已经事先标定好类别。根据是否利用到词汇在各个类别的分布情况，特征加权方法可以分为无监督特征加权方法和有监督特征加权方法。常用的无监督特征加权方法包括TF和TF-IDF等，有监督特征加权方法包括TF-CHI、TF-OR和TF-RF等。现有的特征加权方法主要针对文本分类问题设计，但是视频和文本的词汇分布情况存在较大差异。因此，本文针对词汇在视频中的分布情况，提出了一个新的有监督特征加权方法：TF-PRF。通过实验验证，TF-PRF在ADL和UCF数据集上均取得优于其它特征加权方法的识别准确率。

### 1.3.2 主要贡献

本文的主要贡献包括以下几个方面：

（1）提出了基于显著轨迹的特征提取算法。首先定义了轨迹的显著性，评估人眼对轨迹的视觉关注度。然后，提出两个轨迹显著性的实现方案，包括轨迹的外观显著性和运动显著性，分别用来捕捉静态和动态角度的视觉关注度。对这两个轨迹显著性进行融合，得到轨迹的组合显著性。最后，通过对轨迹几何特征及显著性的分析提取轨迹特征。

（2）将特征加权方法引入视频行为识别问题，并针对该问题提出了一个新的特征加权方法：TF-PRF。TF-PRF属于有监督特征加权方法，根据视觉词汇在正、负类别中的分布赋权。相较于传统的特征加权方法，TF-PRF方法中词汇之间有较好的区分度。

（3）在KTH[6]、ADL[7]和UCF[8]数据集上，对本文方法的实验验证。主要包括基于显著轨迹方法、基于特征加权方法及本文方法与现有方法的对比实验。结果表明，本文提出的方法在UCF上的识别效果与最好的结果接近，而在KTH和ADL两个数据集上的识别准确率均高于其它方法。

## 1.4 本文结构

全文主要分为五章，各章的主要内容如下：

第1章首先叙述了本文的研究背景和研究现状，然后探讨了本文方法的动机和主要的贡献，最后说明了本文的结构安排。

第2章对视频行为识别方法进行综述。根据视频特征的表示形式，将现有的研究方法分为基于全局特征表示、基于局部特征表示和基于层次特征表示的方法，对各类方法的代表性工作进行了概要介绍。

第3章描述了基于显著轨迹和特征加权的视频行为识别方法。对该方法的主要过程进行了详细的阐述，包括：显著轨迹提取、核直方图计算、特征加权及行为分类。在显著轨迹提取过程，首先给出轨迹的显著性定义，包括外观显著性、运动显著性和组合显著性，然后介绍基于显著轨迹的特征提取算法，最后讨论显著轨迹的性质。在核直方图计算过程，根据核直方图的定义介绍视频的运动边界核直方图。在特征加权环节，提出新的特征加权方法TF-PRF，并描述了特征加权算法的过程。该章最后讨论了显著轨迹提取和核直方图计算的时间复杂度。

第4章是本文的实验部分， 实验在KTH、ADL和UCF行为数据集上进行。首先比较了基于显著轨迹方法的性能差异，然后分析了各个特征加权方法的识别准确率随词汇数目的变化情况，最后给出本文方法和现有方法的比较结果，并讨论了算法的时间效率。

第5章给出全文工作的总结并提出对未来工作的展望。

# 第2章 视频行为识别方法综述

过去二十多年里，国内外众多学者就视频行为识别这一问题展开研究并取得令人瞩目的成果。按照特征的表示方式，现有的视频行为识别方法主要可以划分为以下三类：

（1）基于全局特征表示的方法：定位视频中的行为主体并利用全局形状或运动信息进行建模。

（2）基于局部特征表示的方法：利用特征点或轨迹检测算子获取局部区域，并使用特征描述符进行特征表示，该类方法不依赖于行为主体的准确定位。

（3）基于层次特征表示的方法：不孤立地描述局部区域的特征，而是通过层次化的表示形式对局部区域之间的时空关系进行建模。

## 2.1 基于全局特征表示的方法

基于全局特征表示的方法对全局形状或运动信息进行建模。首先检测并定位视频中的行为主体，获取到以行为主体为中心的感兴趣区域（Region of Interest, ROI），然后在ROI所界定的范围内提取特征。基于全局特征表示的方法，按照所使用的特征可以划分为基于形状和基于光流的方法。

### 2.1.1 基于形状的方法

形状（亦称剪影，轮廓）特征广泛应用于早期的视频行为识别方法中。单帧形状对应人体运动过程某一时刻的状态，因此只包含静态信息。多帧形状则描述了人体运动过程中的状态转变，因此可以对行为进行动态建模。基于形状的方法依赖于前景目标的分割，所以该类方法主要应用在监控场景中。

Bobick和Davis[19]运用前景形状为已知的运动创建时序模板。每个时序模板对应着一幅向量图像，其中图像每个点的向量值都由视频序列中相应空间位置的运动特征函数决定。主要的时序模板包含：运动能量图像（Motion Energy Image, MEI）和运动历史图像（Motion History Image, MHI）。其中，MEI反映了运动的范围，而MHI反映了运动的趋势。对未知运动进行识别时，使用相同的方法将其转化为向量图像，并通过和时序模板的距离比较确定运动类型。

Roy等人[20]扩展了Bobick和Davis[19]的工作，提出新的运动表示方法：姿态能量图像（Pose Energy Image, PEI）。该文首先定义人体运动的关键姿态并将每个形状关联到最相似的关键姿态。对于每个关键姿态，其PEI是视频序列中对应该姿态的所有形状所构成的平均图像。为了提高算法效率，该文首先运用姿态统计特征减少候选类别，然后通过PEI特征比较得出最终分类结果。

Veeraraghavan等人[21]利用Kendall形状理论获取人体形状的统计特征，并对人体行为进行参数建模和非参数建模。其中，参数模型主要包括自回归模型（Autoregressive, AR）及自回归滑动平均模型（Autoregressive and Moving Average, ARMA），描述了形状的时域变化。非参数模型则采用了动态时间规整（Dynamic Time Warping, DTW）算法，对模板序列和测试序列进行非线性的时间归一化。

Gorelick等人[2]把人体动作当作时空域上的形状，首先提取时空形状，并通过解泊松方程的方式提取特征，最后比较特征向量之间的距离识别行为类别。Guo等人[22]同样对时空形状进行描述，形状内每个点都可以用一个13维的特征向量来表示，并利用特征向量的协方差矩阵获取时空形状的特征，基于黎曼度量对两个协方差矩阵的相似性进行计算。

Liang等人[23]利用改进后的形状上下文选取形状模板，通过模板匹配将形状序列转换成符号序列，然后使用变长的马尔科夫模型（Variable-Length Markov Model, VLMM）对符号序列进行训练。训练好的VLMM可转换成隐马尔科夫模型（Hidden Markov Model, HMM），对行为进行分类。

当形状的特征向量维度较高时，可以通过流形学习来降低维度。Wang和Suter[24]运用局部保持投影（Locality Preserving Projections, LPP）学习人体行为。LPP是线性流形学习算法，但保留了非线性方法的重要属性，并且更易于计算。Liu和Jia[25]利用有监督的邻域保持嵌入（Neighborhood Preserving Embedding, NPE）来降低形状的特征维度，并且采用隐条件随机场（Hidden Conditional Random Fields, HCRF）对动作的时序关系进行建模。

形状为视频行为识别提供了重要信息。但是，基于形状的方法依赖于静态背景条件或人体分割算法。在背景变化或摄像头移动的情况下，要提取到可靠的前景目标比较困难，此时基于形状的方法不适用。

### 2.1.2 基于光流的方法

另外一类基于全局特征表示的方法使用了光流特征。Polana和Nelson[26]对人体进行跟踪，并把人体运动划分为多个循环过程，对其中各个时空网格计算光流特征描述符。为了补偿人体运动过程中的尺寸变化，该方法对图像尺寸进行动态调整。最后，通过模板匹配的方式，对视频序列的行为进行识别。Efros等人[27]提出了基于光流的时空运动描述符，该方法同样需要对人体进行定位和跟踪。在以人体为中心的时空块中，光流场被分解成水平和垂直分量，每个分量又可进一步分解成正、负通道。该方法通过k-近邻算法，在预先标定类别的视频库中搜索未知行为的类别。

将ROI分割成网格形式，可以描述身体各个部位的特征。Lertniphonphan等人[28]将人体形状分割成多个区域，分别计算各区域的方向光流直方图（Histogram of Oriented Optical Flow, HOOF），并对局部特征向量进行拼接，得到视频帧的全局表示。

为了消除背景/摄像头运动的影响，Mahbub等人[29]将随机抽样一致性算法（Random Sample Consensus, RANSAC）运用在光流场上，过滤背景光流并保留和运动区域相关的部分，实现对人体的准确定位。人体区域进一步被分割成多个块，取各块中特征点数量的变化率作为视频帧的特征向量。

Ali和Shah[30]从光流导出一系列运动特征，包括散度（Divergence）、涡度（Vorticity）、对称场（Symmetric Fields）、非对称场（Asymmetric Fields）等，并对运动特征进行主成分分析（Principal Component Analysis, PCA），得到运动模式。最后，该方法将运动模式与多实例学习（Multiple Instance Learning, MIL）相结合，对行为视频进行分类。

## 2.2 基于局部特征表示的方法

基于局部特征表示的方法将视频序列分成多个局部区域，单独地进行表示。与基于全局特征表示的方法相比，基于局部特征表示的方法更能适应环境等因素的变化，因此，该类方法在过去十年内逐步受到重视。基于局部特征表示的方法通常直接从视频中提取特征，不需要使用人体定位算法。

词袋模型（Bag-of-Visual-Words, BoVW）广泛应用于基于局部特征表示的方法中。对于输入视频，该模型首先提取时空兴趣点（Spatio-Temporal Interest Points, STIP）或运动轨迹，并使用特征描述符表示局部区域。然后，对特征向量进行聚类得到视觉词汇表，将每个特征向量量化到距离最近的词汇当中，进而每个视频都可以表示成词汇的直方图。最后，通过支持向量机（Support Vector Machine, SVM）等分类算法对视频行为进行分类。

### 2.2.1 特征点检测算法

特征点检测算法通常通过最大化特征方程的方式提取到在特定位置和尺度下的时空兴趣点。Laptev和 Lindeberg[9]把Harris算子扩展到时空域，对视频中的每一点计算二阶矩矩阵（Second Moment Matrix），并在以矩阵特征值为参数的拐角函数中，取局部极大值所在位置的点作为检测结果。

在多数场景中，上述特征检测算法提取到的兴趣点过于稀疏，部分对行为识别起到关键作用的特征点甚至被遗漏掉。针对该不足，Dollar等人[10]提出了更为密集的特征点检测算法，在响应函数中使用了时域高斯平滑核以及空域Gabor滤波，并且同样以局部极大值所对应的点作为检测结果。

紧随其后，有更多的特征点检测算法相继被提出[31-33]。Oikonomopoulos等人[31]采用熵值计算的方式提取时空兴趣区域，而Willems等人[32]将Hessian计算方法扩展到时空域，利用3维Hessian矩阵的行列式来检测兴趣点。Lee和Chen[33]构造像素点邻域的直方图表示，并据此形成响应函数来提取兴趣点。这些特征点检测算法都把视频当作一个3维立方体，在时间和空间上使用相同的滤波，因此存在各向同性的空域滤波问题。为了获取到更有效的特征，Shabani等人[34]引入电路表示法，推导出视频信号的时间因果关系多尺度表示。在此基础上，运用各向异性的空域滤波提取时空兴趣点。

在特征点检测算法中，单单依靠响应函数并不能有效地区分前景和背景物体，检测到的兴趣点不可避免会有一部分落在背景区域。为了消除噪声的影响，Chakraborty等人[35]将外围抑制模板运用于检测到的兴趣点，通过非极值抑制方法过滤背景点，并采用时序约束加强过滤效果。该方法检测到的兴趣点更具稳定性和判别性。

### 2.2.2 运动轨迹提取算法

跟踪兴趣点在各帧的位置，可以提取到运动轨迹。Matikainen等人[36]利用KLT跟踪算法[37]得到轨迹片段，并使用BoVW模型表示人体行为。Messing等人[7]同样通过KLT跟踪算法提取运动轨迹，在对数极坐标中对速度历史进行统一量化，然后通过混合马尔科夫链为人体行为建模。上述两个方法虽然都用到同一个轨迹跟踪算法，但前一个方法关注的是短轨迹，而后一个方法关注的是长轨迹，持续跟踪并替换丢失的特征点。

Wang等人[38]提出密集轨迹的概念，按照固定时空间隔对特征点采样，跟踪得到密度较大的轨迹集合。在该方法中，静止或者位移过大的轨迹都被移除。与KLT跟踪算法相比，该方法所提取到的轨迹不仅更加密集，而且更具鲁棒性及可重复性。

和传统的轨迹获取机制不同，Wu等人[39]首先计算光流，然后利用粒子平流过程产生粒子轨迹。粒子轨迹进一步被分解为前景运动和背景运动两部分，前景运动又包括刚体运动和关节运动。其中，背景运动和刚体运动对应着一个低秩矩阵，通过秩优化可以提取到前景物体的运动轨迹。

### 2.2.3 特征描述符

特征描述符用于捕捉时空兴趣点和轨迹邻域内的特征（如外观和运动等）。Dollar等人[10]分别构造亮度、亮度的梯度和光流三种信息的特征描述符，使用特征值拼接，局部直方图和全局直方图三种方法，将时空立方体内所有像素的特征描述转换成一个特征向量，并使用PCA对特征向量进行降维。实验结果表明，识别效果最好的特征向量由梯度和特征值拼接方式获得。

Laptev等人[4]将以兴趣点为中心的时空块划分成网格形式，对每个网格单元分别计算方向梯度直方图（Histograms of Oriented Gradient, HOG）[40]和光流直方图（Histograms of Optical Flow, HOF），以表征局部区域的外观和运动信息，把各个网格单元的直方图归一化并串联得到最终的特征描述符。

与特征检测算子类似，2维的图像描述符可以扩展到3维的视频数据。Scovanner 等人[41]对SIFT描述符[42]进行扩展以获取时空信息，Klaser 等人[43] 将HoG直方图扩展到3维空间，而Willems等人[32]由SURF图像描述符[44]得到扩展的SURF（Extended SURF, ESURF）描述符。上述描述符都采用网格分割的方法，将局部时空块划分成多个子块，分别描述各个子块并串联得到时空块整体的特征向量。进行网格分割方法的原因在于其能够描述区域内部特征的时空分布关系。

Wang等人[38]使用运动边界直方图（Motion Boundary Histogram, MBH）[45]来消除背景运动引起的噪声。MBH首先区分光流场的水平和垂直分量，然后分别计算每个分量的梯度，最后构造直方图对梯度方向进行量化。

### 2.2.4 视觉词汇学习方法

词汇学习的目的是获取紧凑有效的视觉词汇表，增强行为视频之间的区分度，进而提高分类的结果。作为传统的词汇学习方法，k-means聚类算法直接对特征向量分组，由聚类中心构造词汇表。但是k-means算法获得的词汇表不具有语义上的意义。Liu等人[3]对该方法进行改进，在应用k-means得到初始词汇表后，利用信息论方法对词汇表的质量进行评估，并据此更新词汇表。

Kong等人[46]将特征维度的降低过程和视觉词汇的学习过程相结合，提出了加权自适应度量学习算法（Weighted Adaptive Metric Learning, WAML）。WAML算法通过把原始数据映射到类间距离最小的子空间达到降低维度的目的。与PCA等方法不同，WAML算法的降维操作并不在预处理过程执行，而是直接融入到聚类过程。因此能够获得类别信息的反馈。另外，针对传统聚类算法硬量化的缺点，WAML算法将特征向量量化到多个词汇，每个词汇赋予不同的权值，从而获得更精确的行为表示。

K-means算法所获得的视觉词汇往往由出现概率相近的高频特征组成，而Ballan等人[47]则认为在行为分类问题中，中频词汇更具辨别性，并运用基于半径的聚类算法学习视觉词汇。为了减少词汇的不确定性，该方法在聚类过程中引入高斯核函数实现对特征向量的软量化。

Jiang等人[48]提出具有类别一致性的k-means聚类算法（Class Consistent K-Means, CCKM）。CCKM是一种有监督的聚类算法，每个聚类中心都和一个行为类别相对应。对每个行为类别，CCKM都能学习到与之对应的词汇表。算法在计算特征向量和聚类中心的距离时引入了惩罚因子。当两个特征向量对应不同的行为类别时，惩罚因子放大了它们之间的距离，因此CCKM更能区分那些容易混淆的类别。

## 2.3 基于层次特征表示的方法

基于局部特征表示的方法存在条件独立的假设，孤立地对各局部区域进行描述。为了改善这一问题，基于层次特征表示的方法引入分层的思想，通过层次结构来表示低层结构中各局部区域之间的关系。

Kovashka和Grauman[49]利用层次化的组合词汇来描述时空兴趣点邻域之间的关系。算法首先提取局部区域的外观和运动特征，量化成视觉词汇，然后以兴趣点为中心构造邻域内所包含点的组合特征。对于每一层，其组合特征都可以递归地表示高一层的视觉词汇。

Gilbert等人[50]运用关联规则挖掘视觉词汇，在每个行为类别中查找关联度高的特征。算法首先使用2D Harris检测子在3个方向的2维子空间中提取兴趣点，并且以兴趣点为中心，利用网格对邻域内的空间关系进行编码，然后递归地使用Apriori算法查找出现频率高的编码并将其作为复合特征交由上一层处理，最后通过各层挖掘得到的高频特征对人体行为进行表示。

Wang等人[51]提出层次化的语法规则模型来表示人体动作。对于一个动作实例，其语法分析树底层包含动作的基本元素，如时空兴趣点。迭代地在每一层挖掘语法规则，这些语法规则对应高阶的组合特征结构。相较于BoVW模型，该模型能利用高阶的统计特征，因此识别效果较优。

以往的方法普遍使用人为设计的特征对视频进行表示，而Le等人[52]通过独立子空间分析（Independent Subspace Analysis, ISA）直接从视频中学习特征，并构造卷积神经网络模型。在模型的每一层，利用PCA对ISA方法提取到的特征进行降维，最终形成层次化的特征表示。

组件模型（Part-based Model）可以形成特征的层次表示。Wang和Mori[53]将人体分解成多个组件，每个组件和视频帧的局部区域相对应，并使用隐条件随机场（Hidden Conditional Random Field, HCRF）对行为建模。HCRF模型取组件作为隐状态，并结合局部特征和全局特征，以获取充裕的信息量。该文同时通过最大化似然估计和最大化边界两种方式训练模型的参数，其中，最大化边界学习方法更适用于含有复杂隐结构的模型。

Raptis等人[54]同样引入组件模型来描述人体行为，并构造视频的中层表示。对于输入视频，首先提取密集轨迹，计算每对轨迹之间的相似度。然后进行凝聚式聚类获得轨迹组，以轨迹组为中层组件构造人体行为。最终使用图模型描述组件之间的关联结构，通过子图匹配的方式确定行为类别。

在基于层次特征表示的方法中，局部区域之间不存在条件独立的假设，因此该类方法可以获得更完整的行为表示。在基于层次特征表示的方法中，主要的难点来自局部区域之间关联结构的构造。

## 2.4 小结

本章对视频行为识别的研究方法进行综述。按照行为的特征表示方式，主要分为基于全局特征表示、基于局部特征表示和基于层次特征表示的识别方法。在基于全局特征表示的方法中，根据所用的信息，分成基于形状和基于光流的方法进行叙述。在基于局部特征表示的方法中，首先介绍了常用的特征检测算法，包括特征点检测算法和运动轨迹提取算法，然后叙述了主要的特征描述符，最后对视觉词汇的学习方法进行描述。在基于层次特征表示的方法中，对层次结构的构建方式进行介绍。这三类视频行为识别方法的总结可参照表2-1。

**表2-1** 三类视频行为识别方法的比较

|  |  |  |  |
| --- | --- | --- | --- |
|  | 基于全局特征表示的方法 | 基于局部特征表示的方法 | 基于层次特征表示的方法 |
| 概要介绍 | 定位前景区域，对区域整体进行特征表示 | 分别对局部区域进行特征表示 | 由底层语义构造中层语义 |
| 特征表示 | 形状、光流等 | 时空兴趣点、  运动轨迹等 | 层次组合特征 |
| 识别算法 | 模板匹配、DTW、  HMM、HCRF等 | SVM 、KNN等 | HCRF及其它图模型 |

# 第3章 基于显著轨迹和特征加权的视频行为识别方法

本文提出基于显著轨迹和特征加权的视频行为识别方法，该方法主要包括4个过程：显著轨迹提取，核直方图计算，特征加权和行为分类，如图3-1所示。对于输入的行为视频，首先，分析轨迹的显著性，据此提取显著轨迹。其次，使用运动边界核直方图对显著轨迹进行特征描述。第三，通过BoVW将视频转换到视觉词汇的向量空间，利用TF-PRF特征加权方法计算各个视觉词汇的权值。最后，利用基于SVM的分类算法对视频行为分类。本章将按照视频行为识别框架，分别对各个过程的算法进行介绍。

Flow-v2.emf

**图3-1** 基于显著轨迹和特征加权的视频行为识别框架

## 3.1 显著轨迹提取

KLT算法[37]和密集轨迹[38]都考虑了轨迹的几何性质，对拐角点进行跟踪，两者的主要区别在于后者保留的轨迹密度较高。但是，拐角点并非都落在感兴趣的运动区域。当背景存在干扰物体，或者摄像机运动时，检测到的拐角点有部分位于背景区域，此时这两个算法所提取的轨迹存在较大的噪声。

本文针对现有轨迹提取算法存在的噪声问题，提出了基于显著轨迹的特征提取算法。显著性分析用于检测容易引起视觉注意力的区域，而视觉注意区域往往和前景的物体相对应。本文通过对轨迹的显著性分析，提取人眼感兴趣的轨迹，减少背景噪声的影响。

在本节中，首先介绍本文给出的轨迹显著性定义，然后介绍本文提出的基于显著轨迹的特征提取算法，最后分析讨论显著轨迹的性质。

### 3.1.1 轨迹显著性定义

Area.emf

**图3-2** 最大对称外围区域示例（红色矩形和蓝矩形分别表示红色、蓝色中心点的最大对称外围区域）

本文采用中心-外围（Center-Surround）显著性方法来计算轨迹的显著值。对于任意给定的视频，记其第个点所对应的轨迹为，其中，为该点在第帧的坐标向量，而为轨迹的长度。则轨迹在第帧的最大对称外围区域定义如下[12]：

|  |  |  |
| --- | --- | --- |
|  |  | (3-1) |

其中，和为位置偏移量，和分别为视频帧的宽和高。

如图3-2所示, 最大对称外围区域具有自适应性：同一帧上的不同点，其外围区域是不同的。其中，落在边缘位置的点（蓝点）对应的外围区域较小，而落在中间位置的点（红点）对应的外围区域较大。这与常理相符，因为中间位置的点所处的语境较丰富，可进行特征比较的区域也相应较大。

轨迹在单帧的显著值可通过中心点和外围区域的特征值比较得到。直观地，若一个点容易引起视觉注意力，则其自身特征（外观，运动等）和外围区域差异较大。反之，不易引起视觉关注的特征点与外围区域的特征相似。

记为单帧的显著值，取轨迹在各帧的均值作为轨迹的显著值：

|  |  |  |
| --- | --- | --- |
|  |  | (3-2) |

下文将从轨迹的外观、运动以及组合特征实现轨迹的显著性：

**外观显著性：**通过灰度（亮点）图像可以获取物体的外观概况，因此采用灰度特征计算轨迹的外观显著性。记为单帧的外观显著值，本文简化了文献[12]中的计算公式，采用中心点和外围区域的灰度差异计算：

|  |  |  |
| --- | --- | --- |
|  |  | (3-3) |

其中，为经过高斯滤波处理的灰度图像，是外围区域的灰度平均值：

|  |  |  |
| --- | --- | --- |
|  |  | (3-4) |

视频单帧的像素显著值集合可形成一幅显著图。为了获得前后一致的显著值，将每幅显著图上的值都规范化到区间，然后通过各帧的显著性均值计算轨迹的外观显著值：

|  |  |  |
| --- | --- | --- |
|  |  | (3-5) |

其中，是规范化的单帧外观显著值。

**运动显著性：**当视频帧的灰度范围受限时，单纯从静态信息很难区分背景和前景物体。而前景和背景的运动情况存在差异，因此引入动态显著性有助于提高识别效果。记为单帧的运动显著值，主要包括以下三个计算步骤：

Step 1.计算光流以描述视频前后两帧的运动；

Step 2. 对光流图像进行高斯滤波处理，并计算积分直方图；

Step 3. 计算中心点和外围区域光流直方图的卡方距离，得到轨迹单帧的运动显著值：

|  |  |  |
| --- | --- | --- |
|  |  | (3-6) |

其中，为点处，HOF描述符第个bin的值，而为外围区域的光流平均值：

|  |  |  |
| --- | --- | --- |
|  |  | (3-7) |

与轨迹外观显著性一致，将显著图中的值规范化到区间，得到轨迹的运动显著值：

|  |  |  |
| --- | --- | --- |
|  |  | (3-8) |

其中，是规范化的单帧运动显著值。

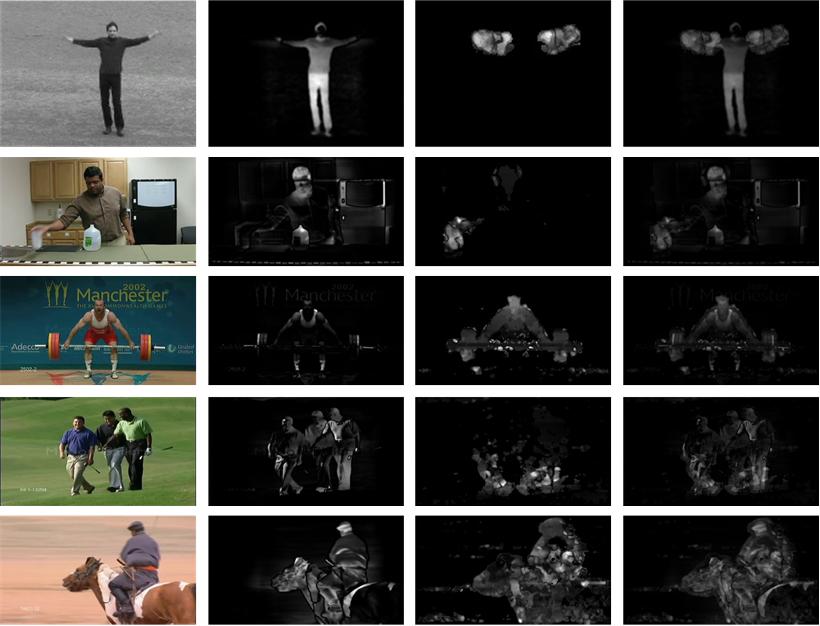
**组合显著性：**依靠单一的显著性并不能检测到足够有效的轨迹，因此本文对外观显著性和运动显著性进行融合，以获取到互补的视觉信息。记为单帧的组合显著值，通过线性组合的方式来计算：

|  |  |  |
| --- | --- | --- |
|  |  | (3-9) |

则轨迹的组合显著值为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-10) |

其中，参数控制着外观显著性和运动显著性的权重。当时，组合显著性退化成为运动显著性，当时，组合显著性则退化成为外观显著性。在本文中，赋予外观显著性和运动显著性相等的权值，即。

****

**图3-3** 显著图比较（从左到右分别为：原始图像，外观显著图，运动显著图和组合显著图）

图3-3展示了本文三种显著性方法的检测效果。其中，组合显著性的定位效果最为准确。外观显著性提取到灰度分布突出的区域，而运动显著性检测到动作显著的区域，组合显著性则融合了两者的检测结果。

### 3.1.2 基于显著轨迹的特征提取算法

本文通过对轨迹的显著性分析，提出了基于显著轨迹的特征提取算法（Salient Trajectory-based Feature Extraction, STFE）。轨迹的显著性评估了其受到人眼关注的程度，而视频行为识别是对人眼关注的行为进行分类，因此，在视频行为识别中，显著值较大的轨迹对分类结果的影响也较大。

本文提出的显著轨迹同时满足两个条件：（1）具有重要的几何特征（如位于拐角处），以易于跟踪特征点；（2）具有较大的显著值，对应着引起视觉注意力的区域。

首先，考虑轨迹的几何特征。为了保证跟踪结果的稳定性和可靠性，所选取的特征点应该落在拐角处或椒盐纹理处。从特征点的几何特征来看，其一阶导数矩阵

|  |  |  |
| --- | --- | --- |
|  |  | (3-11) |

需要满足条件：的最小特征值大于给定阈值[55]。对于不同图像，矩阵的特征值分布并不相同，因此选取的阈值需要具有自适应性。令变量表示矩阵在每个点处的最小特征值，可取最大值的倍数作为自适应阈值。对满足条件的特征点进行跟踪，提取到初始轨迹集。在跟踪过程中，通过光流计算特征点在前后帧的位移[56]。

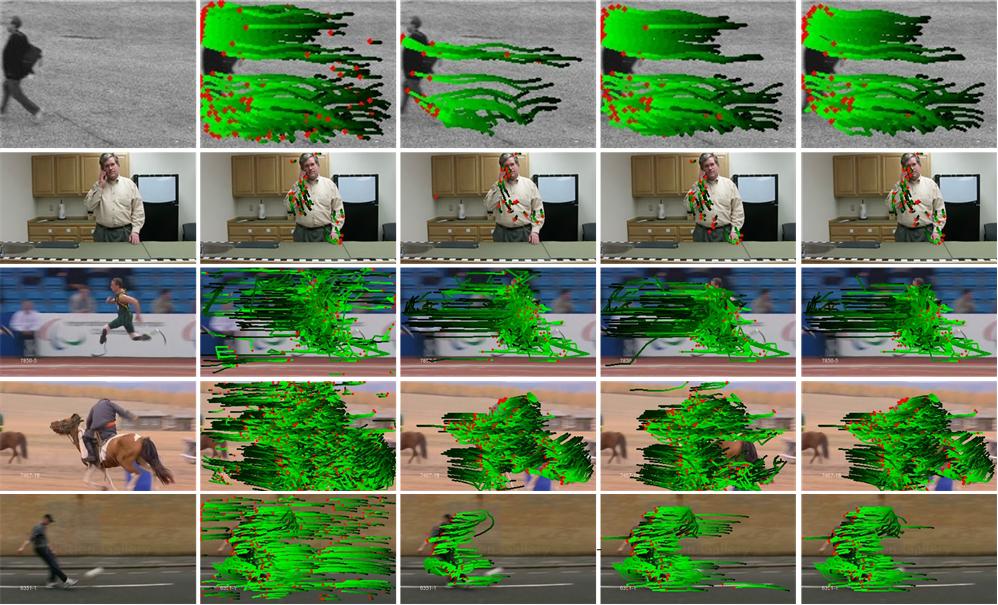
然后，考虑轨迹的显著性。初始轨迹集包含背景噪声，可利用轨迹的显著值进行过滤。在初始轨迹集中，仅保留显著性大于阈值的轨迹，过滤掉不易引起人眼关注的轨迹。阈值的确定过程为：计算每一帧的平均显著值，取在各帧平均值的倍数作为自适应阈值。

图3-4给出了显著轨迹的提取流程图，算法的详细介绍参见算法3-1。

为方便起见，本文将外观显著轨迹、运动显著轨迹及组合显著轨迹分别记为Salient-A、Salient-M和Salient-C。图3-5比较了这三种轨迹，其中， Salient-C的定位效果较优，互补了Salient-A和Salient-M之间漏检或错检的问题。

Trajectory_Flowchart.emf

**图3-4** 显著轨迹提取流程图：(a) 原始视频序列；(b) 初始轨迹集；(c) 外观显著图；(d) 运动显著图；(e) 组合显著图；(f) 显著轨迹



**图3-5** 轨迹集合的比较（从左到右分别为：原始图像，密集轨迹[38]，Salient-A，Salient-M和Salient-C。红点表示当前位置，绿点表示运动历史）

|  |
| --- |
| **算法3-1** 基于显著轨迹的特征提取（STFE） |
| 输入： - 视频  - 初始轨迹密度参数  - 显著轨迹密度参数  - 轨迹跟踪长度  输出： - 显著轨迹集  1：将视频转换到灰度空间  2：在视频初始帧每个点处，由公式(3-11)计算矩阵  3：选取初始轨迹集，其中，是轨迹在初始帧的最小特征值）  4：跟踪初始轨迹集在各帧的位置，并计算各帧的显著值  for  4.1：通过光流计算前后两帧的位移向量，由中值滤波（）得到当前帧的位置  4.2：由公式(3-3)、(3-6)及(3-9)计算当前帧的显著值  4.3：计算当前帧的平均显著值  end  5：计算各条轨迹的显著值  for  由公式(3-10)计算  end  6：提取显著轨迹集 |

### 3.1.3 显著轨迹的性质

**显著轨迹的数学意义：**单帧的显著值等价于对轨迹点的高斯差分滤波[11]：

|  |  |  |
| --- | --- | --- |
|  |  | (3-12) |

其中，和为高斯函数的标准差。为了获取到较大的带宽，取值为无穷大，在实际应用中以最大对称外围区域的特征均值替代该滤波操作。对应窗口尺寸最小的高斯滤波器，可选取尺寸为的二项式滤波器。单帧的显著性等价于高斯差分滤波，而轨迹的显著性则等价于高斯差分滤波的均值：

|  |  |  |
| --- | --- | --- |
|  |  | (3-13) |

其中，为经过二项式滤波的图像，其原始图像可以是灰度图或光流图，是外围区域在中的特征均值。在实际应用中可扩展公式(3-13)，根据所使用的特征选择合适的距离公式。如在灰度图中，选灰度差的绝对值作为距离公式，计算外观显著性。在光流图中，取光流直方图的距离为距离公式，计算运动显著性。

**显著轨迹和密集轨迹[38]的比较**：显著轨迹和密集轨迹都通过考虑轨迹的几何特性来获取初始轨迹集，但两者对初始轨迹集运用不同的过滤规则：（1）密集轨迹分析单条轨迹的性质，首先计算轨迹在各帧位置的标准差，当标准差小于阈值时，认为属于静态轨迹，予以移除，当标准差大于另一阈值时，认为轨迹随机性较大，也将其过滤。此外，在前后两帧中位移较大的轨迹也被当作噪声过滤掉。该过滤规则的一个重要特征是，对运动幅度较大的视频产生的轨迹密度较大，反之也成立。（2）显著轨迹并不仅仅考虑单条轨迹的几何性质，而且由轨迹之间的特征差异计算显著值。另外，阈值设置为平均显著值的倍数，而不是固定的值。因此显著轨迹提取方法更具自适应性，视频的轨迹密度受运动幅度的影响较小。

图3-5比较了密集轨迹和显著轨迹的效果。从图上可以看出，Salient-C能够获取到视觉关注、位于前景运动区域的轨迹。而密集轨迹不能很好地区分前景和背景区域，存在严重噪声，这是因为该方法仅仅考虑了轨迹的几何特征，没有对轨迹的显著性进行分析。

如图3-6所示，当运动幅度较小时，Salient-C也能够提取到静态的轨迹，这是因为该方法同时考虑了静态和动态两个方面的显著性，只要其中一个方面的特征突出，轨迹都能被检测到。而密集轨迹方法提取到的轨迹密度很小，这是因为此时人体区域的轨迹被当作静态轨迹过滤掉。



**图3-6** 密集轨迹与组合显著轨迹在运动幅度较小时的比较（从左到右分别为：原始图像，密集轨迹和Salient-C）

结合对图3-5和图3-6的分析可知，相较于密集轨迹，Salient-C的密度不易受到行为主体运动幅度的影响。当运动幅度较大时，Salient-C能够过滤相对稳定的背景运动，而当运动幅度较小时，Salient-C能够通过对静态特征的分析获得足够的轨迹。轨迹提取方法的稳定性对视频行为的识别具有重要意义。

## 3.2 核直方图计算

局部直方图作为高效的特征描述符，广泛应用于视觉识别领域。然而，传统直方图存在硬量化的问题（每个方向仅仅量化到最相近的一个bin中，其它bin的取值都为0），量化过程的粗糙性造成信息的丢失。为了进行更准确的特征表示，本文采用核直方图描述显著轨迹。

在本节中，首先描述核直方图的定义及本文提出的核直方图查表法，然后介绍基于运动边界核直方图的特征表示过程。

### 3.2.1 核直方图定义

核直方图是传统直方图的平滑量化版本。令和分别表示点处某一特征（梯度，光流等）的大小和方向，则对于一个维度为的传统直方图，点在第个bin处的分量为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-14) |

其中，是四舍五入取整函数。

对于同样维度的核直方图，则有：

|  |  |  |
| --- | --- | --- |
|  |  | (3-15) |

其中，和是方向特征向量， 为第个bin的主方向，为高斯平滑函数的参数，本文取[57]。在核直方图中，点在每个bin中都有分量值，分量值的大小与两者的方向相似度有关，方向越相近，该bin所对应的分量也越大。

核直方图中存在大量的指数运算，造成计算开销过大。为了提高算法效率，本文提出了核直方图查表法（Kernel Histogram Table Look-up, KHTL），如算法3-2所示。虽然核直方图查表法中存在方向的量化，但划分的区间足够小（），能够准确地表示特征点。对于所有的特征点，表都固定不变，因此只需计算一次，减少了算法的计算时间。

|  |
| --- |
| **算法3-2**核直方图查表法（KHTL） |
| 输入： - 特征点  输出： - 特征点的核直方图  1：将划分为个足够小的区间（），每个区间的方向特征向量为（）  2：计算大小为的矩阵，矩阵各个元素  3：在个区间中选取方向最相近的一个逼近，计算核直方图。记选中的区间为，则有 |

### 3.2.2 运动边界核直方图

本文取MBH的核平滑版本：运动边界核直方图（Motion Boundary Kernel Histogram, MBKH）作为轨迹的特征描述符。同时，在轨迹显著性的计算过程中，选择光流核直方图（Kernel Histograms of Optical Flow, KHOF）代替HOF。KHOF和MBKH都可以利用公式(3-15)计算得到，两者的主要区别在于计算过程中所用到的特征。其中，KHOF使用了光流特征，而MBKH使用了运动边界特征。MBKH实质上是光流的梯度核直方图，表示了特征点之间的相对运动，因此能够对摄像头的运动起到抑制效果。

首先，分解光流场的水平分量和垂直分量，以便独立计算两个分量的MBKH。在上，梯度方向对应着水平运动变化幅度最大的方向，而梯度大小表示水平运动的变化速率，上梯度的含义同理可推。

然后，以轨迹为中心构造时空立方体，并进行网格划分。立方体的长度与轨迹的长度相等，立方体的横截面是大小为32×32的正方形。为了获取到时空结构信息，将立方体划分为单元数目为的时空网格，计算每个网格单元的MBKH并拼接得到最终的特征向量。为了获取到任意矩形区域的核直方图，本文引入了积分直方图[58]。在积分直方图的构造过程中，使用KHTL算法（算法3-2）计算各点的特征值。基于MBKH的特征表示（MBKH-based Feature Representation，MBFR）过程如算法3-3所示。

|  |
| --- |
| **算法3-3** 基于MBKH的特征表示（MBFR） |
| 输入： - 显著轨迹  - 水平、垂直及时间维度划分的单元数  输出： - 轨迹的MBKH，其中和分别是水平、垂直分量的MBKH  1：计算光流，分解光流的水平、垂直分量得到和  2：计算和的梯度积分直方图  3：以轨迹上的点为中心构造时空立方体  4：将划分为单元数为的时空网格  5：对上的每个网格单元分别计算和，并连接各单元的MBKH得到最终的特征向量  初始化和为空集  for  5.1：通过积分直方图得到和  5.2：，  end |

综上讨论可知，轨迹的MBKH向量维度为：分量数×网格单元数×bin数。其中，方向分量数目为2，网格参数的取值分别为，，bin的数目为8，因此特征向量的维度是192。

按照BoVW模型，对轨迹的MBKH特征向量聚类。本文利用k-means算法得到个聚类中心，构造视觉词汇表。每个特征向量都通过欧氏距离的比较，量化到最近的词汇中去。因此，每个视频在向量空间中都可以表示成词频向量，其中表示词汇在视频中的出现频率。

## 3.3 特征加权

**图3-7** 文本数据集（20 Newsgroups）与视频数据集（KTH）的词汇分布比较（*N*：视频总数，*A*：包含词汇的视频数目，*B*：词汇在各个视频的出现频率之和，出现概率=*A/N*，平均出现频率=*B/A*）

特征加权常常用于文本分类问题以改善分类效果。但是，现有的视频行为识别方法并没有对特征加权在该问题中的作用进行系统化的研究。另外，文本的特征加权方法是否适用于视频行为识别也有待验证。如图3-7所示，视频和文本的词汇分布情况存在较大差异。图3-7左侧表明，KTH视频数据集中词汇的出现概率远高于20 Newsgroups文本数据集[[2]](#footnote-2)。另外，由图3-7右侧可知，20 Newsgroups中大多数词汇在一个文本中只出现一次，因此传统特征加权方法通常忽略词汇在文本中的多次出现。然而，KTH中词汇在单个视频中的出现频率并不相同，可以利用该信息增强词汇的区分度。

本文针对视频和文本的词汇分布差异，提出了TF-PRF（Term Frequency-Pseudo Relevance Frequency）特征加权方法。在本节中，首先给出TF-PRF权值的定义，然后介绍视频的TF-PRF特征加权算法。

### 3.3.1 TF-PRF权值定义

本文提出的TF-PRF特征加权方法是对TF-VRF（Variant of Relevance Frequency）的改进。TF-VRF和TF-PRF都属于有监督特征加权方法。对于有监督特征加权方法，词汇在各个类别中的权值不同，记为词汇在类别中的权值，则词汇在视频中的权值为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-16) |

其中为类别数目。

在TF-VRF方法中，因子定义为[18]：

|  |  |  |
| --- | --- | --- |
|  |  | (3-17) |

其中，是正类别中包含词汇的视频数目，而是负类别中包含词汇的视频数目。当计算词汇在类别中的权值时，为正类别，而剩余类别的集合为负类别，标记为。

本文对TF-VRF方法的改进主要包括两个方面：（1）使用词汇在各视频的出现频率之和代替包含该词汇的视频数目；（2）引入类别数目作为调整因子来规范化词汇的出现频率。改进后的特征加权方法称为TF-PRF，在TF-PRF中，因子为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-18) |

其中，是词汇在正类别中各视频的出现频率之和，而是词汇在负类别中各视频的出现频率之和。由图3-7可知，一个词汇在单个视频中可能多次出现，因此（）。对于任意词汇，当且仅当其在所有包含该词汇的视频中只出现一次，等号成立。

**表3-1** TF-VRF与TF-PRF方法的权值比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 词汇 | *n*11 | *n*01 | *f*11 | *f*01 | |*C*| | *FW(k,c)* | |
| TF-VRF | TF-PRF |
| *k*1 | 5 | 5 | 25 | 5 | 2 | 1.0000 | 1.8184 |
| *k*2 | 5 | 5 | 5 | 25 | 2 | 1.0000 | 0.5499 |
| *k*3 | 5 | 5 | 25 | 5 | 6 | 1.0000 | 4.7004 |

表3-1显示，TF-PRF能根据词汇的分布情况给出更合理的权值。词汇、和的分布情况存在差异，其中，出现在5个正类别视频和5个负类别视频中，在每个正类别视频的平均出现频率为5，在每个负类别视频的平均出现频率为1。同样出现在5个正类别视频和5个负类别视频中，但是其在正、负类别每个视频的平均出现频率分别为1和5，与的情况正好相反。的分布与相似，唯一的差别是所在的数据集包含6个类别，负类别可划分为5个子类，每个子类中只有一个视频包含，而所属的数据集包含2个类别，负类别不可再划分。在合理的情况下，三个词汇的权值关系为。但是，在TF-VRF方法中，三个词汇的因子取值相等，而TF-PRF能根据词汇分布的差异，给出合适的权值。



**图3-8** TF-VRF与TF-PRF在各行为视频数据集的权值分布比较

图3-8进一步比较了TF-VRF和TF-PRF在KTH、ADL和UCF行为视频数据集上的权值分布情况。可见，TF-VRF方法的权值过度集中（谷峰值较大，且分布曲线狭长），因此词汇之间的区分度较小。而TF-PRF方法的权值较分散，尤其是在ADL和UCF上。因此，相较于TF-VRF，TF-PRF方法中的词汇有更好的区分度，有助于视频行为的分类。

### 3.3.2 TF-PRF特征加权算法

利用本文提出的TF-PRF特征加权算法，可以计算视觉词汇在各个视频中的权值。首先利用训练数据统计词汇在各个类别的出现频率，然后根据出现频率计算，最后在每个视频中计算词汇的TF-PRF权值。详细的计算过程如算法3-4所示。

|  |
| --- |
| **算法3-4** TF-PRF特征加权（TF-PRF） |
| 输入： - 训练数据集中各个视频的词汇频率向量，其中，为词汇在视频中的出现频率  - 测试视频的词汇频率向量  输出： - 训练数据集的词汇特征向量，其中，为词汇在视频中的权值  - 测试视频的词汇特征向量  1：定义辅助矩阵，大小为，初始化矩阵元素为0  2：统计词汇在训练数据集各类别中的出现频率  for  读取视频的类别  for      end  end  3：根据词汇在正、负类别的分布计算PRF因子  for  for      根据公式(3-18)计算  end  end  4：计算各个视频（包括训练集及测试视频）中TF-PRF的权值  for  for  根据公式(3-16)计算  end  根据公式(3-16)计算  end |

通过TF-PRF算法得到各个词汇的权值后，需对特征向量进行归一化处理：

|  |  |  |
| --- | --- | --- |
|  |  | (3-19) |

## 3.4 行为分类

本文使用带核的SVM作为分类器。与多通道特征组合方法的不同之处在于，本文只使用了MBKH这一单通道特征。对于单通道特征，SVM的核函数定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | (3-20 ) |

其中，和分别为视频、中词汇的权值向量，为训练视频之间距离的平均值。

如算法3-5所示，基于SVM的行为分类（SVM-based Action Classification，SVMAC）算法首先计算训练视频之间的核函数值，训练SVM。然后计算未标定视频与训练视频之间的核函数，根据训练好的SVM判断视频行为。

|  |
| --- |
| **算法3-5** 基于SVM的行为分类（SVMAC） |
| 输入： - 训练视频集的词汇特征向量  - 测试视频的词汇特征向量  输出： - 测试视频的类别标签  1：计算训练集中每一对视频的卡方距离  2：计算训练集的平均卡方距离  3：由公式(3-20)计算训练集中每一对训练视频的核函数  4：由训练视频间的核函数训练SVM  5：计算测试视频和各个训练视频之间的核函数  for  由公式(3-20)计算  end  6：由SVM对步骤5的结果进行判断，得到分类结果 |

## 3.5 时间复杂度

针对视频行为识别问题，本文的主要改进工作包括：显著轨迹提取，核直方图计算和特征加权。其中，特征加权的执行效率远远快于前两个过程，因此，本文只讨论显著轨迹提取和核直方图计算的时间复杂度，为方便起见，将这两个过程合称为特征表示。

**显著轨迹提取的时间复杂度：**计算轨迹的显著性时，视频每帧的主要操作包括光流计算、图像高斯滤波和显著图计算：（1）光流计算的时间复杂度为[56]。其中，和分别为视频帧的宽和高，而是复合参数，对应在每个像素处进行的操作数。（2）高斯滤波操作的时间复杂度为。高斯滤波通过模板和像素的卷积操作得到滤波后的像素值。针对离散数据，本文采用了二项式滤波模板，分别从、方向对图像进行平滑处理。（3）显著图计算的时间复杂度为。主要的操作为像素和外围区域的特征比较，在每个像素上该操作需要常数时间。

**核直方图计算的时间复杂度：**在核直方图计算过程，主要的操作是积分直方图的计算。积分直方图统计从图像起始位置到各个像素位置所围矩形区域的直方图信息，在每一点处所需的操作数由直方图的bin数决定。MBKH中bin的数目固定为8，因此积分直方图计算的时间复杂度为。

综上讨论可知，视频一帧操作的时间复杂度为。因此，对于一个帧的视频，特征表示所需的时间复杂度为，该时间复杂度主要取决于光流计算操作。

## 3.6 小结

本章提出了基于显著轨迹和特征加权的视频行为识别方法。该方法主要包括4个过程：第一，显著轨迹提取。定义了轨迹的显著性，分别从静态和动态视觉注意力的角度实现了外观显著性和运动显著性，基于两者的组合提取显著轨迹。第二，核直方图计算。对传统直方图进行平滑量化得到核直方图，通过MBKH描述显著轨迹的特征。第三，特征加权。提出TF-PRF特征加权方法，由视觉词汇在正、负类别视频中的分布情况计算权值。第四，行为分类。利用非线性SVM对行为进行学习和分类。

# 第4章 实验结果与分析

本文的实验在单机环境下进行，主要硬件配置为Intel Pentium D 3.00GHz CPU和1.5G内存，操作系统为Linux，编程语言为C/C++，使用了OpenCV和Libsvm[59]的相关API。

在实验中，核直方图计算及行为分类过程分别采用MBFR算法（算法3-3）及SVMAC算法（算法3-5）。对于显著轨迹提取过程，比较了STFE算法（算法3-1）在不同阈值下的实验结果。对于特征加权过程，比较了本文提出的TF-PRF算法（算法3-4）及传统特征加权方法的识别结果。

## 4.1 实验数据

本文实验采用了业内著名的行为视频基准数据集：KTH[6]、ADL[7]和UCF[8]，表4-1列出了这三个数据集的统计情况，图4-1至图4-3分别展示了三个数据集的样本。

**表4-1** KTH、ADL和UCF数据集总结

|  |  |  |  |
| --- | --- | --- | --- |
|  | KTH[[3]](#footnote-3) | ADL[[4]](#footnote-4) | UCF[[5]](#footnote-5) |
| 数据来源 | Schuldt等人[6] | Messing 等人[7] | Rodriguez 等人[8] |
| 年份 | 2004 | 2009 | 2008 |
| 视频数量 | 599 | 150 | 150 |
| 行为类型数量 | 6 | 10 | 10 |
| 行为主体数量 | 25 | 5 | - |
| 场景 | 室内+室外 | 室内 | 室内+室外 |
| 摄像头运动情况 | 轻微运动 | 静止 | 运动 |
| 分辨率 | 160×120 | 1280×720 | 720×480 |
| 数据集描述 | 简单人体动作 | 日常生活行为 | 自然场景运动视频 |

**评价标准：**本文在三个数据集上均使用Leave-One-Out（LOO）交叉验证。对于KTH和ADL，每一轮取一个行为主体的所有视频作为验证集，其他行为主体的视频作为训练集，循环执行直到所有主体的数据都得到验证。对于UCF，为了增加数据样本，添加视频的水平翻转版本[38]。每一轮都取一个视频作为验证数据（此时从训练集中移除该视频的水平翻转版本），其余的视频作为训练集。对视频行为识别结果的评价采用通用的标准，即所有行为类别的平均准确率：

|  |  |  |
| --- | --- | --- |
|  |  | (4-1) |

其中，是类别的准确率。

**参数设置：**，，与密集轨迹[38]的默认参数值保持一致。，默认的全词汇数目。，均衡显著轨迹的查全率和查准率。

**其它设置：**为了加快算法的执行速度，在实验中将ADL数据集的分辨率降低为原先的1/5，UCF数据集的分辨率降低为原先的1/2，而KTH数据集的分辨率保持不变。

KTH.emf

**图4-1** KTH数据集的行为类别示例

**ADL.emf**

**图4-2** ADL数据集的行为类别示例

**UCF.emf**

**图4-3** UCF数据集的行为类别示例

## 4.2 实验结果

### 4.2.1 基于显著轨迹方法的结果比较

本小节对本文提出的三种基于显著轨迹的方法进行比较，包括基于Salient-A、基于Salient-M以及基于Salient-C的方法，这三个方法分别对应STFE算法中参数取值为1、0及0.5时的情形。选取Wang等人[38]提出的基于密集轨迹的方法[[6]](#footnote-6)作为比较基准。本部分实验主要在ADL和UCF两个数据集上进行，特征加权方法固定为TF（本文取词频的对数函数值代替词频，参见公式4-2）。

如表4-2所示，基于Salient-C的方法取得最优的识别准确率。该方法在ADL和UCF上的平均准确率分别比基于密集轨迹的方法高6%和4.5%左右，这说明了对轨迹的显著性分析能够提取到有效的轨迹。另外，基于Salient-A和基于Salient-M的方法单独使用时识别效果都要比组合后的方法差，这表明了多特征组合在显著性分析中的作用。

**表4-2** 基于密集轨迹与基于显著轨迹方法的平均准确率（%）比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 密集轨迹[38] | Salient-A  STFE（） | Salient-M  STFE（） | Salient-C  STFE（） |
| ADL | 92.00 | 94.67 | 91.33 | **98.00** |
| UCF | 85.35 | 88.31 | 84.38 | **90.08** |

图4-4比较了基于密集轨迹和基于Salient-C的方法在ADL上的混淆矩阵。基于Salient-C的方法能够正确识别大多数行为，而基于密集轨迹的方法存在着多个误判。除了“UseSilverware”行为之外，基于Salient-C的方法在其它行为的准确率都不低于基于密集轨迹的方法。其中，识别效果改进最大的是“AnswerPhone”行为，提高了40%，这是因为“AnswerPhone”与“DialPhone”这两个行为的轨迹很相似，因此容易误判，而基于Salient-C的方法能利用两个场景中物体的不同显著性来区分行为。

图4-5分别展示了基于密集轨迹和基于Salient-C的方法在UCF上的混淆矩阵。在两种方法中，最容易误判的行为都来自“Skate”，准确率分别为50%和58%。基于Salient-C的方法在一半的行为上取得100%的准确率，并且在除了“Run”之外的其它行为上都取得不低于基于密集轨迹方法的准确率。在识别效果上，改进最大的行为是“Kick”，其准确率提高了15%。

表4-3比较了密集轨迹和Salient-C在ADL和UCF上的密度。从表中可以看出，Salient-C的密度比密集轨迹稳定。密集轨迹和Salient-C在UCF上的密度都比ADL数据集大，这是因为UCF的场景复杂，运动幅度大。由两个数据集的密度均值比可知，Salient-C在自然场景和监控场景下的轨迹密度差异小于密集轨迹。另外，由轨迹密度的标准差可知，Salient-C在单个数据集内部的轨迹密度差异同样小于密集轨迹。综上所述，在两种轨迹中，Salient-C受到运动幅度的影响较小，提取到的轨迹密度较稳定。造成这一现象的原因是密集轨迹和Salient-C定义了不同的过滤规则。

**表4-3** 密集轨迹与Salient-C的密度比较（视频的轨迹密度=每15帧内的平均轨迹数目）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 密集轨迹[38] | | Salient-C | |
| ADL | UCF | ADL | UCF |
| 最大密度 | 434.27 | 7097.14 | 404.87 | 4422.62 |
| 最小密度 | 2.93 | 6.21 | 216.05 | 198.31 |
| 密度均值 | 136.91 | 1981.20 | 321.13 | 1149.88 |
| 密度标准差 | 81.14 | 1535.93 | 34.50 | 661.52 |
| 密度均值比  （UCF/ADL） | 14.47 | | 3.58 | |

**图4-4** 基于密集轨迹与基于Salient-C的方法在ADL数据集上的混淆矩阵



**图4-5** 基于密集轨迹与基于Salient-C的方法在UCF数据集上的混淆矩阵

总结本小节的实验结果，得出结论：（1）基于Salient-C的方法识别效果优于基于密集轨迹的方法以及两类基于单特征显著轨迹的方法；（2）基于Salient-C的方法较不易受到视频行为主体运动幅度的影响，提取到的轨迹密度比基于密集轨迹的方法稳定。

### 4.2.2 基于特征加权方法的结果比较

本小节对多个基于特征加权方法的识别效果进行分析，包括两个无监督特征加权方法TF和TF-IDF，五个有监督特征加权方法TF-CHI、TF-OR、TF-RF、TF-VRF以及本文所提出的TF-PRF。为了研究各个特征方法在不同词汇数目下的表现，使用特征选取方法来筛选词汇。具体地，以各个特征加权方法除去TF的权值作为特征选择的权值（对TF-IDF来说为IDF因子，对有监督特征加权方法来说为因子）。由于TF方法不包含特征选取因子，因此只保留其全词汇形式。本部分实验在三个数据集上进行，显著轨迹固定为Salient-C。

两个无监督特征加权方法定义如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| （1）TF： | |  | (4-2) | |
| （2）TF-IDF： |  | | | (4-3) |

其中，是视频的总数，是包含词汇的视频数目。

有监督特征加权方法的定义如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| （1）TF-CHI： |  | | | | | (4-4) |
| （2）TF-OR： | |  | | | (4-5) | |
| （3）TF-RF[17]： | | |  | (4-6) | | |

其中，是正类别中包含词汇的视频数目，是正类别中没有包含词汇的视频数目，是负类别中包含词汇的视频数目，而是负类别中没有包含词汇的视频数目。TF-VRF和TF-PRF则分别由公式(3-17)、(3-18)定义。

图4-6描述了各个特征加权方法在KTH上的平均准确率与词汇数目的关系。随着词汇数目的增加，各个特征加权方法的平均准确率呈现出提高的趋势。词汇数目对准确率的影响较小，如当词汇数目大于1750时，所有特征加权方法的平均准确率都达到96.00%以上。而在词汇数目为250时，除TF-IDF外其它特征加权方法的平均准确率都达到93.00%以上。在该数据集上，各个特征加权方法的性能差异不大，并没有一个特征加权方法表现出绝对的优势。



**图4-6** 特征加权方法在KTH数据集上平均准确率与词汇数目的关系

表4-4列出各个特征加权方法在KTH上的最优平均准确率。可以看出，最高和最低的准确率相差很小，保持在0.5%以内。本文提出的TF-PRF平均准确率达到97%，对应的词汇数为2500，这与TF方法在全词汇下的准确率相同。

**表4-4** 特征加权方法在KTH数据集上的最优平均准确率（%）比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | TF | TF-IDF | TF-CHI | TF-OR | TF-RF | TF-VRF | TF-PRF |
| 准确率 | 97.00 | **97.17** | 96.83 | **97.17** | 97.16 | 97.00 | 97.00 |
| 词汇数目 | 4000 | 3250 | 1750 | 3250 | 2500 | 2500 | 2500 |

Salient-C + TF-PRF.emf

**图4-7** Salient-C + TF-PRF在KTH数据集上的混淆矩阵

图4-7展示了词汇数目为2500时Salient-C+TF-PRF方法的混淆矩阵。从图上可以看出，该方法的误判主要来自于行为“Jogging”和“Running”，这是合理的，因为这两个行为很相似，即使依靠人眼进行判断都容易混淆。

图4-8描述了各个特征加权方法在ADL上平均准确率与词汇数目的关系。随着词汇数目的增加，各个特征加权方法的平均准确率呈现出不同的变化趋势。其中，TF-IDF、TF-RF、TF-VRF及TF-PRF的平均准确率都随着词汇数目的增加而提高，TF-CHI的平均准确率随着词汇数目的增加而降低，而TF-OR则呈现先降后提的趋势。这说明词汇数目对于不同特征加权方法的影响是不同的。另外，当词汇数目大于2500时，TF-PRF的识别结果好于其它的特征加权方法。



**图4-8** 特征加权方法在ADL数据集上平均准确率与词汇数目的关系

表4-5对各个特征加权方法在ADL上的最优平均准确率进行比较。只有TF-PRF特征加权方法能够获得完全准确的识别结果。此时，词汇数目大于等于3500。次优的结果98.67%分别由TF-CHI及TF-VRF在词汇数目为1500及4000时获得，而TF方法在全词汇下的平均准确率为98.00%，与其余特征加权方法的最优结果相等。

**表4-5** 特征加权方法在ADL数据集上的最优平均准确率（%）比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | TF | TF-IDF | TF-CHI | TF-OR | TF-RF | TF-VRF | TF-PRF |
| 准确率 | 98.00 | 98.00 | 98.67 | 98.00 | 98.00 | 98.67 | **100.00** |
| 词汇数目 | 4000 | 3000 | 1500 | 1500 | 4000 | 4000 | 3500 |

图4-9给出Salient-C+TF-PRF方法在词汇数目为3500时的混淆矩阵。由于该方法的准确率为100%，此时混淆矩阵对角线的元素都为100，其它位置的元素都为0。

Salient-C + TF-PRF.emf

**图4-9** Salient-C + TF-PRF在ADL数据集上的混淆矩阵



**图4-10** 特征加权方法在UCF数据集上平均准确率与词汇数目的关系

图4-10描述了各个特征加权方法在UCF上平均准确率随词汇数目的变化情况。各个特征加权方法受词汇数目的影响并不相同，其中，TF-CHI和TF-OR两种方法的表现比较稳定，受词汇数目的影响很小，而其它特征加权方法的平均准确率都随着词汇数目的增加呈现出提高的趋势。另外，当词汇数目大于等于3500时，TF-PRF方法的识别结果优于其它的特征加权方法。

表4-6展示了各个特征加权方法在UCF上的最优平均准确率。最优的两个平均准确率90.91%、90.08%分别由TF-PRF和TF特征加权方法在全词汇下获得。虽然TF-CHI和TF-OR的表现比较稳定，但两者的最优平均准确率在所有特征加权方法中是最低的。

**表4-6** 特征加权方法在UCF数据集上的最优平均准确率（%）比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | TF | TF-IDF | TF-CHI | TF-OR | TF-RF | TF-VRF | TF-PRF |
| 准确率 | 90.08 | 88.79 | 87.98 | 87.98 | 89.31 | 89.31 | **90.91** |
| 词汇数目 | 4000 | 4000 | 3000 | 2500 | 4000 | 4000 | 4000 |

Salient-C + TF-PRF.emf

**图4-11** Salient-C + TF-PRF在UCF数据集上的混淆矩阵

图4-11所给出Salient-C+TF-PRF方法在全词汇下的混淆矩阵，从中可以发现，最易误判的行为主要来自于“Skate”和“Walk”。

总结各个特征加权方法在三个数据集上的实验结果，得出结论：（1）词汇数目对各个特征加权方法的影响并不相同。除了TF-CHI和TF-OR方法外，其它特征加权方法在三个数据集上的平均准确率都随词汇数目的增加呈现提高趋势；（2）无监督特征加权方法的表现并不劣于有监督特征加权方法，如TF方法在三个数据集上都有不错的识别效果；（3）在KTH上，各个特征加权方法有相近的平均准确率，而在ADL和UCF上，最优的识别结果均由本文提出的TF-PRF方法获得。

### 4.2.3 与现有方法的结果比较

本小节对本文方法与现有方法的结果进行比较，所选取的比较方法均采用LOO验证方式。

**表4-7** KTH数据集上的平均准确率（%）比较

|  |  |  |
| --- | --- | --- |
| 方法 | 准确率 | 特征表示类型 |
| **Salient-C + TF-PRF** | **97.00** | 局部特征 |
| **Salient-C + TF** | **97.00** |
| Gilbert等人[50] | 95.7 | 层次特征 |
| Chakraborty等人[35] | 96.35 | 局部特征 |
| Raptis和Soatto[60] | 94.5 |
| Liu等人[3] | 93.8 |
| Wang和Zhao[14] | 91.3 |
| Kong等人[46] | 88.81 |
| Rapantzikos等人[13] | 88.30 |
| Dollar等人[10] | 81.17 |
| Liu和Yuen[61] | 81.5 | 全局特征 |

如表4-7所示，在KTH上，本文提出的Salient-C+TF-PRF、Salient-C+TF均取得高于其它方法的识别准确率。具体地，Salient-C+TF-PRF的平均准确率比Gilbert等人[50]基于层次特征表示的方法高1.3%，比Liu和Yuen[61]基于全局表示的方法高15.5%。考虑同一特征表示类型，Salient-C+TF-PRF比两类基于显著特征点的方法分别高8.7%[13] 、5.7% [14]，这是因为本文方法关注轨迹静态和动态两方面的互补信息，更具鲁棒性。

**表4-8** ADL数据集上的平均准确率（%）比较

|  |  |  |
| --- | --- | --- |
| 方法 | 准确率 | 特征表示类型 |
| **Salient-C + TF-PRF** | **100.00** | 局部特征 |
| **Salient-C + TF** | **98.00** |
| Benabbas等人[62] | 81 | 层次特征 |
| Wang等人[63] | 96 | 局部特征 |
| Messing等人[7] | 89 |
| Raptis和Soatto[60] | 82.67 |
| Satkin和Hebert[64] | 80 |
| Matikainen等人[65] | 70 |

表4-8对ADL上的结果进行比较。从表中可以看出，本文提出的Salient-C+TF-PRF方法取得最优的平均准确率100%。本文方法的准确率比目前的最优结果[63]高4%，这说明轨迹的显著性分析有助于提取到有效的特征。该数据集的结果验证了，Salient-C+TF-PRF方法不仅适用于简单的人体动作，而且对于复杂的生活行为同样有效。

如表4-9所示，在UCF上，本文的Salient-C+TF-PRF方法取得与最好方法相近的识别准确率90.91%，该结果比Kovashka和Grauman[49]基于层次特征表示的方法高3.64%，比Lan等人[67]基于全局特征表示的方法高7.21%。最优的结果由Shabani等人[34]基于局部特征表示的方法获得，这是因为该方法运用了改良的时空特征检测方法，提取到高质量的特征。

**表4-9** UCF数据集上的平均准确率（%）比较

|  |  |  |
| --- | --- | --- |
| 方法 | 准确率 | 特征表示类型 |
| **Salient-C + TF-PRF** | **90.91** | 局部特征 |
| **Salient-C + TF** | **90.08** |
| Kovashka和Grauman[49] | 87.27 | 层次特征 |
| Le等人[52] | 86.5 |
| Shabani等人[34] | 91.5 | 局部特征 |
| Wang等人[38] | 88.2 |
| Wang等人[66] | 85.6 |
| Derpanis等人[67] | 81.5 |
| Lan等人[68] | 83.7 | 全局特征 |

### 4.2.4 算法的时间效率

由3.5的讨论可知，视频特征表示的时间复杂度为。在实验过程中，对于一个帧数为100，空间分辨率为360×240的视频，本文方法进行特征表示所需的时间约为160秒。作为比较，在基于中层表示的方法中[54]，同样帧数的视频所需的处理时间为200秒，并且不包含光流的计算时间。可知本文的算法的时间效率相对较好。

## 4.3 小结

本章在KTH、ADL和UCF数据集上对基于显著轨迹和特征加权的视频行为识别方法进行实验验证。在基于显著轨迹方法的结果比较中，分析了三种基于显著轨迹方法的性能，并以基于密集轨迹的方法作为比较基准，实验结果验证了本文提出的基于Salient-C方法的优越性。在基于特征加权方法的对比实验中，比较了各个特征加权方法在不同数量视觉词汇下的识别效果，结果表明，在ADL和UCF数据集上，本文提出的TF-PRF方法均能取得优于其它特征加权方法的性能。另外，通过本文方法和其它现有方法的比较得知，本文提出的Salient-C+TF-PRF方法在UCF上取得与最优方法相近的结果90.91%，在KTH和ADL上分别取得最优的结果97.00%和100%。最后，对算法时间效率的验证表明，基于显著轨迹和特征加权的方法不仅取得较好的识别效果，而且有较好的时间效率。

# 第5章 总结与展望

## 5.1 全文总结

视频行为识别是计算机视觉领域的热点问题，本文对该课题进行研究，综述了目前的主要方法，并提出了基于显著轨迹和特征加权的视频行为识别方法，该方法的主要过程包括显著轨迹提取、核直方图计算、TF-PRF特征加权和SVM行为分类。本文主要的研究工作和贡献包括：

（1）对视频行为识别方法的综述。阅读了视频行为识别领域内的大量文献，按照特征表示的方式，将现有方法划分为基于全局特征表示、基于局部特征表示和基于层次特征表示的方法，分别对各类方法进行介绍。

（2）提出了基于显著轨迹和特征加权的视频行为识别方法。首先定义了轨迹的显著性，并分别从静态和动态角度定义了轨迹显著性的两种实现：外观显著性和运动显著性，基于两类显著性的线性组合提取显著轨迹。另外，本文将特征加权方法引入到视频行为识别领域，并提出了一个新的有监督特征加权方法：TF-PRF，针对视觉词汇在各个类别中的出现频率计算词汇的权值。

（3）在KTH、ADL和UCF数据集上对本文方法的实验验证。首先比较了三种基于显著轨迹方法的识别准确率，然后分析了主要的无监督特征加权方法（TF、TF-IDF）和有监督特征加权方法（TF-CHI、TF-OR、TF-RF和TF-VRF）在不同的词汇数目下的平均准确率，最后验证了算法的时间效率。通过与现有方法的比较实验可知，本文提出的基于显著轨迹和特征加权的识别方法在UCF上取得与目前最优方法相近的结果，而在KTH和ADL上的结果比目前公开发表的最好结果略优。

## 5.2 工作展望

基于显著轨迹和特征加权的方法在视频行为识别方面取得了较好的效果，但是该方法还存在可改进的空间：

（1）轨迹显著性的研究。为了提高算法的时间效率，本文提出的轨迹显著性定义相对比较简单，并且在显著性的计算过程中主要使用了灰度和光流信息。更为有效的显著性定义以及更多视觉特征的运用有待进一步的研究。

（2）层次特征表示方式的研究。本文主要通过底层特征表示视频行为，而人体行为属于高层语义，怎样在这两层之间引入中层语义，如何定义中层语义的结构，中层语义如何促进高层语义的理解等问题都是有待探索的。

（3）图像行为识别问题的研究。当场景从视频变成图像，视频行为识别问题转变成图像行为识别问题。此时问题空间由三维变到二维，由于图像中不包含动态信息，该问题更具挑战性。怎样调整现有的视频行为识别方法，使其对图像的行为识别问题有效，同样值得关注。

# 参考文献

[1] Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes [J]. Computer Vision and Image Understanding, 2006, 104 (2-3): 249-257.

[2] Gorelick L, Blank M, Shechtman E, et al. Actions as space-time shapes [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2247-2253.

[3] Liu J, Luo J, Shah M. Recognizing realistic actions from videos “in the wild” [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami, USA, 2009: 1996-2003.

[4] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, USA, 2008: 1-8.

[5] Ryoo M S, Aggarwal J K. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities [DB/OL]. http://cvrc.ece.utexas.edu/SDHA2010/Human\_Interaction.html, 2010.

[6] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local svm approach [C]// Proc 17th International Conference on Pattern Recognition (ICPR), Cambridge, UK, 2004: 32-36.

[7] Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints [C]// Proc 12th IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 2009: 104-111.

[8] Rodriguez M D, Ahmed J, Shah M. Action MACH: a spatio-temporal maximum average correlation height filter for action recognition [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, USA, 2008: 1-8.

[9] Laptev I, Lindeberg T. Space-time interest points [C]// Proc Ninth IEEE International Conference on Computer Vision (ICCV), Nice, France, 2003: 432-439.

[10] Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features [C]// Proc Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), Beijing, China, 2005: 65-72.

[11] Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami, USA, 2009: 1597-1604.

[12] Achanta R, Susstrunk S. Saliency detection using maximum symmetric surround [C]// Proc 17th IEEE International Conference on Image Processing (ICIP), Hong Kong, China, 2010: 2653-2656.

[13] Rapantzikos K, Avrithis Y, Kollias S. Dense saliency-based spatiotemporal feature points for action recognition [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami, USA, 2009: 1454-1461.

[14] Wang L, Zhao D. Recognizing actions using salient features [C]// Proc 13th IEEE International Workshop on Multimedia Signal Processing (MMSP), Hangzhou, China, 2011: 1-6.

[15] Vig E, Dorr M, Cox D D. Saliency-based selection of sparse descriptors for action recognition [C]// Proc 19th IEEE International Conference on Image Processing (ICIP), Florida, USA, 2012: 1405-1408.

[16] Debole F, Sebastiani F. Supervised term weighting for automated text categorization [C]// Proc 18th Annual ACM Symposium on Applied Computing (SAC), Melbourne, USA, 2003: 784-788.

[17] Lan M, Tan C L, Su J, et al. Supervised and traditional term weighting methods for automatic text categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4): 721-735.

[18] Quan X, Liu W, Qiu B. Term weighting schemes for question categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 1009-1021.

[19] 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.

[20] Roy A, Sural S, Mukherjee J. Gait recognition using pose kinematics and pose energy image [J]. Signal Processing, 2012, 92(3): 780-792.

[21] Veeraraghavan A, Roy-Chowdhury A K, Chellappa R. Matching shape sequences in video with applications in human movement analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1896-1909.

[22] Guo K, Ishwar P, Konrad J. Action recognition in video by covariance matching of silhouette tunnels [C]// Proc XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), Rio de Janeiro, Brazil, 2009: 299-306.

[23] Liang Y M, Shih S W, Shih A C C, et al. Learning atomic human actions using variable-length Markov models [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(1): 268-280.

[24] Wang L, Suter D. Learning and matching of dynamic shape manifolds for human action recognition [J]. IEEE Transactions on Image Processing, 2007, 16(6): 1646-1661.

[25] Liu F, Jia Y. Human action recognition using manifold learning and hidden conditional random fields [C]// Proc 9th International Conference for Young Computer Scientists (ICYCS), Hunan, China, 2008: 693-698.

[26] Polana R, Nelson R. Low level recognition of human motion [C]// Proc IEEE Workshop on Motion of Non-Rigid and Articulated Objects (MNRAO), Texas, USA, 1994: 77-82.

[27] Efros A A, Berg A C, Mori G, et al. Recognizing action at a distance [C]// Proc Ninth IEEE International Conference on Computer Vision (ICCV), Nice, France, 2003: 726-733.

[28] Lertniphonphan K, Aramvith S, Chalidabhongse T H. Human action recognition using direction histograms of optical flow [C]// Proc 11th International Symposium on Communications and Information Technologies (ISCIT), Hangzhou, China, 2011: 574-579.

[29] Mahbub U, Imtiaz H, Ahad M A R. An optical flow based approach for action recognition [C]// Proc 14th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2011: 646-651.

[30] Ali S, Shah M. Human action recognition in videos using kinematic features and multiple instance learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(2): 288-303.

[31] Oikonomopoulos A, Patras I, Pantic M. Spatiotemporal salient points for visual recognition of human actions [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2006, 36(3): 710-719.

[32] Willems G, Tuytelaars T, Gool L J V. An efficient dense and scale-invariant spatio-temporal interest point detector [C]// Proc 10th European Conference on Computer Vision (ECCV), Marseille, France, 2008: 650-663.

[33] Lee W T, Chen H T. Histogram-based interest point detectors [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami, USA, 2009: 1590-1596.

[34] Shabani A H, Clausi D A, Zelek J S. Improved spatio-temporal salient feature detection for action recognition [C]// Proc 22nd British Machine Vision Conference (BMVC), Dundee, UK, 2011.

[35] Chakraborty B, Holte M B, Moeslund T B, et al. Selective spatio-temporal interest points [J]. Computer Vision and Image Understanding, 2012, 116(3): 396-410.

[36] Matikainen P, Hebert M, Sukthankar R. Trajectons: action recognition through the motion analysis of tracked features [C]// Proc 12th IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Kyoto, Japan, 2009: 514-521.

[37] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision [C]// Proc 7th International Joint Conference on Artificial Intelligence (IJCAI), Vancouver, Canada, 1981: 674-679.

[38] Wang H, Klaser A, Schmid C, et al. Action recognition by dense trajectories [C]// Proc 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA, 2011: 3169-3176.

[39] Wu S, Oreifej O, Shah M. Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories [C]// Proc 13th IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 2011: 1419-1426.

[40] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, USA, 2005: 886-893.

[41] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition [C]// Proc 15th International Conference on Multimedia (MULTIMEDIA), Augsburg, Germany, 2007: 357-360.

[42] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[43] Klaser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3d-gradients [C]// Proc 19th British Machine Vision Conference (BMVC), Leeds, UK, 2008.

[44] Bay H, Tuytelaars T, Gool L V. SURF: speeded up robust features [C]// Proc Ninth European Conference on Computer Vision (ECCV), Graz, Austria, 2006: 404-417.

[45] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance [C]// Proc Ninth European Conference on Computer Vision (ECCV), Graz, Austria, 2006: 428-441.

[46] Kong Y, Zhang X, Hu W, et al. Adaptive learning codebook for action recognition [J]. Pattern Recognition Letters, 2011, 32(8): 1178-1186.

[47] Ballan L, Bertini M, Bimbo A D, et al. Effective codebooks for human action representation and classification in unconstrained videos [J]. IEEE Transactions on Multimedia, 2012, 14(4): 1234-1245.

[48] Jiang Z, Lin Z, Davis L S. Class consistent k-means: application to face and action recognition [J]. Computer Vision and Image Understanding, 2012, 116(6): 730-741.

[49] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition [C]// Proc 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, USA, 2010: 2046-2053.

[50] Gilbert A, Illingworth J, Bowden R. Action recognition using mined hierarchical compound features [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 883-897.

[51] Wang L, Wang Y, Gao W. Mining layered grammar rules for action recognition [J]. International Journal of Computer Vision, 2011, 93(2): 162-182.

[52] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis [C]// Proc 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA, 2011: 3361-3368.

[53] Wang Y, Mori G. Hidden part models for human action recognition: probabilistic versus max margin [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(7): 1310-1323.

[54] Raptis M, Kokkinos I, Soatto S. Discovering discriminative action parts from mid-level video representations [C]// Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, USA, 2012: 1242-1249.

[55] Shi J, Tomasi C. Good features to track [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 1994: 593-600.

[56] Farneback G. Two-frame motion estimation based on polynomial expansion [C]// Proc 13th Scandinavian Conference on Image Analysis (SCIA), Halmstad, Sweden, 2003: 363-370.

[57] Bo L, Ren X, Fox D. Kernel descriptors for visual recognition [C]// Proc 24th Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 2010: 244-252.

[58] Porikli F. Integral histogram: a fast way to extract histograms in Cartesian spaces [C]// Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, USA, 2005: 829-836.

[59] Chang C C, Lin C J. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3):27:1-27:27.

[60] Raptis M, Soatto S. Tracklet descriptors for action modeling and video analysis [C]// Proc 11th European Conference on Computer Vision (ECCV), Heraklion, Greece, 2010: 577-590.

[61] Liu C, Yuen P C. Human action recognition using boosted EigenActions [J]. Image and Vision Computing, 2010, 28 (5): 825-835.

[62] Benabbas Y, Lablack A, Ihaddadene N, et al. Action recognition using direction models of motion [C]// Proc 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010: 4295-4298.

[63] Wang J, Chen Z, Wu Y. Action recognition with multiscale spatio-temporal contexts [C]// Proc 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA, 2011: 3185-3192.

[64] Satkin S, Hebert M. Modeling the temporal extent of actions [C]// Proc 11th European Conference on Computer Vision (ECCV), Heraklion, Greece, 2010: 536-548.

[65] Matikainen P, Hebert M, Sukthankar R. Representing pairwise spatial and temporal relations for action recognition [C]// Proc 11th European Conference on Computer Vision (ECCV), Heraklion, Greece, 2010: 508-521.

[66] Wang H, Ullah M M, Klaser A, et al. Evaluation of local spatio-temporal features for action recognition [C]// Proc British Machine Vision Conference (BMVC), London, UK, 2009.

[67] Derpanis K G, Sizintsev M, Cannons K J, et al. Action spotting and recognition based on a spatiotemporal orientation analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(3): 527-540.

[68] Lan T, Wang Y, Mori G. Discriminative figure-centric models for joint action localization and recognition [C]// Proc 13th IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 2011: 2003-2010.

# 攻读硕士学位期间与硕士学位论文相关的研究成果清单

[1] Yang Yi, **Yikun Lin (林乙坤)**. Human action recognition using integrated model [C]// Proc 2013 5th International Conference on Digital Image Processing (ICDIP 2013). Washington: SPIE Press. (已录用，待发表) (与硕士学位论文第二章相关)

[2] Yang Yi, **Yikun Lin (林乙坤)**. Human action recognition with salient trajectories [J]. Signal Processing. (已录用，待发表) (与硕士学位论文第三章和第四章相关)

# 致 谢

盛年不重来，一日难再晨。三年多的研究生生活即将成为过去，在此，要特别感谢研究生阶段在学术和生活上帮助我的老师、家人和朋友们。我坚信这个阶段的付出终将在之后的工作生活中得到收获。

感谢我的导师衣杨老师，在研究课题的选取、研究方法的把握以及论文的写作等方面给予的指导。感谢师兄师姐们在实验室例会上精彩的展示，让我对本领域内的前沿研究有了更进一步的了解。

感谢我的家人，特别是我的父母，对我生活上点滴的关心，并解决了我经济上的后顾之忧。

感谢10计A的同学们，在生活和学习上的互相帮助。近朱者赤，与众多优秀的人相处更激励了个人的进步。

最后，还要感谢视频行为识别这一课题的众多学者们，本论文正是基于前人的研究才得以完成。

1. 数据来源：www.iresearch.cn [↑](#footnote-ref-1)
2. 数据集链接：http://web.ist.utl.pt/~acardoso/datasets/ [↑](#footnote-ref-2)
3. 数据集链接：http://www.nada.kth.se/cvap/actions/ [↑](#footnote-ref-3)
4. 数据集链接：http://www.cs.rochester.edu/~rmessing/uradl/ [↑](#footnote-ref-4)
5. 数据集链接：http://www.cs.ucf.edu/vision/public\_html/data.html [↑](#footnote-ref-5)
6. 代码链接：http://lear.inrialpes.fr/people/wang/dense\_trajectories [↑](#footnote-ref-6)