西安电子科技大学研究生学位论文
撰写要求（2015年修订版）

作者姓名 张三

指导教师姓名、职称 李四 教授

申请学位类别 工学硕士

学校代码 10701　　学　号 1101110071
分 类 号 TN82　　密　级 秘密

西安电子科技大学

硕士学位论文

西安电子科技大学研究生学位论文撰写要求（2015年修订版）

作者姓名：张三

一级学科：电子科学与技术

二级学科：电磁场与微波技术

学位类别：工学硕士

指导教师姓名、职称：李四 教授

学　　院：电子工程学院

提交日期：20xx年x月

Thesis/Dissertation Guide for Postgraduates of XIDIAN UNIVERSITY

A Thesis submitted to

XIDIAN UNIVERSITY

in partial fulfillment of the requirements

for the degree of Master

in Electromagnetic Field and Microwave Technology

By

Zhang San

Supervisor: Li Si Professor

February 2015

西安电子科技大学
学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同事对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文若有不实之处，本人承担一切法律责任。

本人签名：_____　　日　　期：_____

西安电子科技大学
关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权属于西安电子科技大学。学校有权保留送交论文的复印件，允许查阅、借阅论文；学校可以公布论文的全部或部分内容，允许采用影印、缩印或其它复制手段保存论文。同时本人保证，结合学位论文研究成果完成的论文、发明专利等成果，署名单位为西安电子科技大学。

保密的学位论文在_____年解密后适用本授权书。

本人签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘要

摘要是学位论文的内容不加注释和评论的简短陈述，简明扼要陈述学位论文的研究目的、内容、方法、成果和结论，重点突出学位论文的创造性成果和观点。摘要包括中文摘要和英文摘要，硕士学位论文中文摘要字数一般为 1000 字左右，博士学位论文中文摘要字数一般为 1500 字左右。英文摘要内容与中文摘要内容保持一致，翻译力求简明精准。摘要的正文下方需注明论文的关键词，关键词一般为 3 ～ 8 个，关键词和关键词之间用逗号并空一格。

中文摘要格式要求为：宋体小四、两端对齐、首行缩进 2 字符，行距为固定值 20 磅，段落间距为段前 0 磅，段后 0 磅。

英文摘要格式要求为：Times New Roman、小四、两端对齐、首行不缩进，行距为固定值 20 磅，段落间距为段前 0 磅，段后 0 磅，段与段之间空一行。

关键词：XXX，XXX，XXX，XXX，XXX

ABSTRACT

The Abstract is a brief description of a thesis or dissertation without notes or comments. It represents concisely the research purpose, content, method, result and conclusion of the thesis or dissertation with emphasis on its innovative findings and perspectives. The Abstract Part consists of both the Chinese abstract and the English abstract. The Chinese abstract should have the length of approximately 1000 Chinese characters for a master thesis and 1500 for a Ph.D. dissertation. The English abstract should be consistent with the Chinese one in content. The keywords of a thesis or dissertation should be listed below the main body of the abstract, separated by commas and a space. The number of the keywords is typically 3 to 5.

The format of the Chinese Abstract is what follows: Song Ti, Small 4, justified, 2 characters indented in the first line, line spacing at a fixed value of 20 pounds, and paragraph spacing section at 0 pound.

The format of the English Abstract is what follows: Times New Roman, Small 4, justified, not indented in the first line, line spacing at a fixed value of 20 pounds, and paragraph spacing section at 0 pound with a blank line between paragraphs.

Keywords: XXX, XXX, XXX, XXX, XXX

插图索引

表格索引

符号对照表

符号	符号名称
C	数据库中的任务/类别个数
$\mathbf{W}$	分类矩阵
N_c	第 c 个类别中的样本个数
$\mathbf{x}_{ci}$	来自第c个类别的第i个样本的特征向量
$\mathbf{y}_{ci}$	来自第c个类别的第i个样本的标签
$\mathbf{w}_c$	第c个类别对应的分类参数向量
$\cdots$	

缩略语对照表

缩略语	英文全称	中文对照
BoVW	Bag of Visual Words	词袋模型
DTW	Dynamic Time Warping	动态时间变形
RMTFL	Robust Multi-Task Feature Learning	鲁棒的多任务特征学习
SVM	Support Vector Machines	支持向量机

...

目录

第一章　绪论

1.1　研究背景及意义

随着在过去二十余年中计算机技术的蓬勃发展，人们逐渐希望计算机可以拥有与人类类似的视觉系统，从而可以智能地处理图片和视频序列。基于这个美好的设想，计算机视觉（Computer Vision），一门由统计学、信息论、模式识别、计算机科学等多个领域交叉而成的新兴学科应运而生。人体行为识别[1][2]是当前计算机视觉领域的一个研究热点，顾名思义，即让计算机可以通过算法识别图像或者视频中的人做出的各种行为。随着 Youtube 等视频网站的盛行和视频拍摄器材的普及，越来越多的国内外学者将研究的重点转向了基于视频的人体行为识别技术。

人体行为识别技术具有广阔的应用市场和极大的理论研究价值。在应用方面，它可应用于智能人机交互、视频监控[3]、视频搜索等领域。将人机交互技术与人体行为识别相结合将颠覆传统的人机交互方式，人们再不需要使用鼠标和键盘与电脑交互，而只用对着屏幕做动作就可以完成操作，这对于残疾人和老人等人群十分友好。事实上，这样的技术已经被初步实现并投入了市场。例如微软公司于 2010 年上市的体感外设 Kinect，它通过利用红外摄影结构光编码对于人体的各个结构进行跟踪和识别；在视频监控方面，近年来国内外暴力恐怖事件频发，如新疆七五事件和昆明火车站暴力事件。如果通过计算机分析和识别摄像头传来的视频序列，可以实时监控可疑分子，抑制违法事件的发生；在视频检索方面，准确恰当的视频关键字是视频检索结果正确的前提，而人工的标注不仅仅工作量巨大，而且存在不确定性。通过计算机对视频的内容进行分析识别，使标注视频关键字的工作变的方便且可信，从而也提高了视频检索的准确度。在理论方面，国内外的许多科研机构和高校都已开展了大量关于人体行为识别的项目，在世界顶级计算机视觉会议 CVPR、ICCV 和 ECCV 上，每年都有大量的关于行为识别的研究成果被展示。由上述的例子可知，人体行为识别技术的研究意义重大，有很好的商业前景和理论价值。

通过国内外学者在行为识别领域的辛勤研究，目前已经提出了很多人体行为识别方法。这些方法多数是先提取可以描述行为轨迹的特征，利用特征训练分类器来完成识别。绝大多数行为识别方法都主要将研究的视频样本局限于在可控环境下录制的视频，例如 Weizmann dataset 数据库[4]。这导致了这些方法无法处理充斥着大量噪音的真实视频，从而无法满足工业界的需求。尤其近年来随着视频数据的数量飞速增长，对视频中的行为进行准确的识别已成为亟需解决的问题，因此提出切实可行的人体行为识别方法迫在眉睫。行为识别问题的关键挑战在于利用有限的真实视

频样本并克服光照和遮挡等噪音，提高识别准确度和泛化能力，为视频监控、虚拟现实、人机智能交互等应用提供技术支持，这也是本文研究的重点。

1.2 行为识别的研究现状

随着互联网的飞速发展，特别是 Youtube 、优酷土豆等视频社交网站的普及和视频拍摄成本的降低，使得网络上的视频数量呈爆炸式的增长，且推动了基于视频的行为识别技术的发展。当前国内外主流的行为识别方法可以大致分为顺序模型、基于人体结构的层次模型和基于时空特征的模型[1]等。顺序模型[5]将输入的视频视为一系列的图片帧，将图片帧的特征按时间顺序串联成特征向量来表示行为。由于视频的长度和行为速度的变化，此类方法需要借助于动态时间变形（Dynamic Time Warping, DTW）[6]来对齐视频。顺序模型可以完整的体现特征之间的时序信息，适用于复杂的行为的探测和识别，然而单纯的将特征串联带来了维数灾难等问题；基于人体结构的层次模型将视频中的行为分解成人体各个部位的子运动，从而对人体的各个部位进行建模，1992 年 Chen 和 Lee 提出的 stick 模型[7]将人体的各个部位看做线段，线段之间的角度表示了人体各项运动的自由度，随后基于矩形单元和椭圆单元的模型也相继被提出，基于人体结构的层次模型的主要优点在于可以识别具有复杂结构的行为，此外可以根据人体的机构引入一些先验的约束；基于时空特征的模型利用视频独有的时间信息，通过提取视频中可以描述人体运动轨迹的时空特征，如光流、时空兴趣点对行为进行建模。目前常用的时空特征主要有时空块（Space-Time Volumes），时空轨迹（Space-Time Trajectories）, 时空局部特征（Space-Time Local Features），下文将主要介绍基于这三种特征的行为识别模型。

1.2.1 现有技术分析

（ 1）基于时空视频块的行为识别

基于时空视频块的行为识别先将视频序列划分为一系列的视频块，然后通过衡量视频块之间的相似度来完成视频中行为的识别。为了准确的描述视频块之间的相似度，近年来国内外学者提出了众多的视频块表示和识别方法。Bobick 和 Davis[8]于 2001 年提出通过忽略视频中的背景，只提取包含人体动作的前景来跟踪现状的变化，Ke 等人[9]于 2008 年提出可以自动的挖掘一个行为对应的 3D 视频块的方法，Rodriguez 等人[10]为了使视频块的匹配度更高更方便，提出了一种可以捕捉视频块特征的滤波器。Shechtman 和 Irani[11] 通过从 3D 视频块中估计运动流来识别人体的动作，他们在每个视频块的位置附近采样包含了局部运动流的一部分的视频时空块，通过计算测试样本时空块与给定模板时空块之间的相似度从而可以实现匹配和识别。基于时空视频块的行为识别的主要缺点在于无法识别当多个人同时出现在同一个场

景的行为。许多的方法利用滑动窗口来解决这个问题，然而这又需要大量的计算来进行精确的行为跟踪和定位。此外它在识别那些在空间上无法被分割的行为时效果很差。

（2）基于时空轨迹的行为识别

基于时空轨迹的行为识别将行为视为时空轨迹的集合，人被表示为各个关节点的 2 维或者 3 维集合。当人做一个动作时，关节点在时空上的位置变化形成了轨迹。早期运动轨迹被 Yilmaz 和 Shah 等人[12]直接被用来表示和识别行为，Sheikh 等人[13]将人体表示为 13 个关节点的运动轨迹，他们利用仿射投影算法正规化运动轨迹，从而可以比较在不同视角下录制的视频。Rao 和 Shah[14] 在 2011 年利用皮肤像素检测跟踪人体关键点得到运动轨迹，将行为表示成运动轨迹的峰值位置和间隔的集合，且证明了这种表示方法也具有视角不变性。Heng Wang 等人[15]不再仅仅局限于关注关键点产生的运动轨迹，而是在视频中通过稠密地采样关键点得到密集轨迹，他在密集轨迹的基础上，通过降低摄像机移动抖动的影响提出了目前视频识别和人体行为识别方面效率最高，识别精度最高的特征之一–改进的密集轨迹[16]。基于时空轨迹的行为识别模型可以分析人体行为的细节，而且绝大部分是不受视角影响的。这些方法都建立在可以准确地估计场景中人体关键点位置的基础上，然而人体部位的探测和跟踪目前仍是一个悬而未决的问题，有很多学者正致力于该领域的研究。

（3）基于时空局部特征的行为识别

基于时空局部特征的行为识别模型，顾名思义，即从 3D 时空块中提取局部特征来表示和识别行为。早期的方法是在每一个图片帧上提取局部特征，将它们按时间的顺序串联来描述全局的动作。Chomat 和 Crowley[17] 首先提出了利用 Gabor 滤波器来提取可以描述运动信息的局部特征。Zelnik-Manor 和 Irani[18] 通过在多时间尺度上利用时空特征，从而解决了不同视频之间行为速度不同的问题。另一方面 Laptev 和 Lindeberg[19] 等人将以往用于物体识别的具有尺度不变性的 Harris 滤波器扩展到视频的特征提取上，通过在 3D 视频块中提取稀疏的时空兴趣点来表示行为。Dollar 等人[20]专门为提取具有局部周期运动的局部兴趣点设计了一种滤波器，一旦某个兴趣点被探测到，他们的算法就会在这个点的邻域中提取局部特征。因为 3D 视频块实际上可以视为一个严格的 3D 物体，如果可以提取合适的特征来描述这个物体，那么行为识别问题就可以转化为物体识别问题（Object recognition），受到物体识别方法的启发，研究者们发现将词袋技术(Bag of Visual Words, BoVW)[21]与时空局部特征结合会很好地提升行为识别的效果，具体流程如图 1.1 所示。和物体识别的过程类似，首先从时空块中提取局部特征，预处理之后通过词袋技术结合特征之间的时间与空间信息，最后利用特征训练分类器。

词袋模型通过选择不同的聚类和编码方式产生了很多的变种，其中利用高斯混

图 1.1　基于词袋模型的行为识别方法的主要流程.

合模型得到的 Fisher Vector[22][23][24] 在行为识别领域取得了很好的效果。首先通过原始时空特征得到由若干个高斯函数组成的高斯混合模型，然而将每个高斯函数的方差和协方差的导数串联即可得到对应的 Fisher Vector 。所以 Fisher Vector 的每一段连续维度的特征都对应着一个特定的高斯函数。和传统的 k-means 聚类的词袋模型相比，Fisher Vector 可以描述视频更加高阶的特性，它在各个数据库上的表现证明了它是目前最好的特征之一，它也是本文所用的特征。

1.2.2　现存主要问题

由上文可知，现有的行为识别技术可以较好的识别录制于简单场景下的行为，但由于人体本身的复杂性和行为的多变性使得现有方法的效率和精度都没有达到相关行业的实用要求。所以就整体而言，人体行为识别目前还面临着巨大的挑战。以下是对该领域面临的主要难点的简单阐述。

（1）识别对象的差异性

识别对象的差异性体现在两个方面，分别是行为的差异性和人体的差异性。前者表现为速度、幅度和方向等方面的差异性。后者不但在身高、体形和服装方面存在差异性，而且不同的人实施的同一动作也不可能完全相同，例如一个青年和一个老年同时进行走路这一行为，必定存在速度和步长等差异性。人体动作发生的时间点是不可知的，动作的持续时间也是变化的，这需要算法对动作在时域范围内进行细致的分析和研究，极大的影响了识别精度和计算效率的提升。

（2）环境与背景的多变性

不同于录制于实验室环境下的视频，互联网上的绝大多数视频场景都充斥着遮挡、光照和复杂场景等噪音，这增加了人体定位和跟踪的难度。尽管可以通过滤波等预处理手段对原始特征去噪，但是效果有限。通过深度摄像头得到的深度特征可以避免光照等噪音的影响，但其本身也有精度等方面的劣势。环境带来的影响还有多视角的问题。相机在不同视角拍摄的同一行为在视觉上会产生极大的差异。近年来有学者提出了利用多摄像机融合通过三维重建来实现人体关节点从二维空间到三维空间的映射，从而克服二维空间中由视角变化产生的视觉差异，然而这个方法不仅仅成本昂贵，且三维重建目前也是计算机视觉领域一个未解决的难题。

（3）视频中动作的分割

图 1.2　数据库中的行为类别可以被分为不同的组，在同一组内的行为分享相似的特征信息。

录制于真实环境下的视频中除了要识别的对象之外还包含很多其他的内容。目前行为识别领域的主要研究都直接省略了对象探测这一步，直接从只包含识别对象的视频样本中提取特征，这种方法的实用性很差。所以准确识别真实环境下的视频对象的先决条件是先分割出只含有识别对象的视频序列。然而由于人体行为的复杂度和高自由度，动作分割的难度较大。

1.3　论文的主要研究内容及工作安排

大多数现有的行为识别技术都忽略了行为类别之间的关系，把每个行为当做一个单独的任务来训练分类器。不仅仅如此，本文通过观察，发现数据库中的行为类别存在潜在的结构信息，通过测量类别之间的相似度并聚类，可以将数据库分为若干组，每个组内相似的行为类别可以共享特征信息，如图 1.2 所示。更具体的，可以将每个行为类别视为一个任务，通过基于组结构约束的多任务学习框架可以鼓励类别之间共享信息，从而有效的利用有限的真实视频样本，并在一定程度上缓解噪音问题。在此背景下，如何找到类别之间的相似性度量从而得到数据库的潜在结构，如何在多任务学习的基础上利用这个结构使同一组内的行为类别共享特征信息，从而提高行为识别技术的准确度和泛化能力，正是本文的研究重点所在。

本文首先提出一种基于 SVM 间隔的多任务行为识别方法。众所周知，SVM 间隔可以用来衡量两个类之间的相似程度，如果 SVM 间隔越小说明这两个类越相似，反之亦然。本文对数据库中的每对行为类别分别训练一个 1-VS-1 的 SVM，用支持向量之间的距离来衡量这两个类别之间的相似程度。将所有类别的相似度矩阵聚类

之后可以得到数据库中行为类别的分组结构，最后利用基于组结构约束的多任务学习框架训练出每一类行为对应的分类器。

其次本文又提出了一种基于互信息的多任务行为识别方法。我们通过观察发现高斯混合模型中不同的高斯函数都捕捉了对象的不同的特性和运动模式，不同的高斯函数与不同的行为类别的关系紧密程度也不尽相同。相似的类别更倾向于和同一个高斯函数有更紧密的关系。如图 1.2 所示，踢和踢球这个行为十分相似，那么它们就会分享由同一个高斯函数捕捉的特征信息。于是我们选择直接从特征本身下手，通过计算每一个高斯函数对于不同类别的互信息作为类别之间的相似度度量。与前一种方法相比，既然多任务学习的目的是鼓励相似类别之间分享特征信息，那么直接从特征本身的角度计算类别之间的相似性当然也更加的直接和易于理解，进而提高行为识别的性能。

论文的具体章节安排如下：

第一章：介绍了人体行为识别技术的研究背景和意义、目前行为识别技术的研究现状和主要技术难点和本文的主要研究内容和章节安排。

第二章：介绍了多任务学习方法的发展，主要介绍了几种常用的多任务学习方法。

第三章：提出了一种基于 SVM 间隔的多任务行为识别方法，主要是 SVM 间隔的计算，目标函数的建立及优化求解，通过实验证明了该方法的可靠性。

第四章：提出了一种基于互信息的多任务行为识别方法，通过计算高斯函数的互信息来得到数据库的潜在结构，并最后给出了实验结果和分析。

第五章：首先对论文的研究工作作出总结，在此基础上展望了未来的工作方向。

第二章　多任务学习方法

2.1　引言

传统的机器学习方法中大多是把一个对象看做一个任务，每次只研究一个任务，这些方法称为单任务学习。例如在行为识别领域中，把每一类行为看做任务，每次仅对一个任务单独学习其对应的分类器。这种方法的缺点在于它忽略了隐藏在训练样本之间的丰富的可用信息，且在实际应用中，由于人力和经济等多种因素的制约，我们可以得到的训练样本是十分有限的，在这种训练样本不充分的情况下运用单任务学习得到的模型很可能存在过拟合等问题。通过观察可以发现，不仅仅是在数据库中，在现实生活中要学习的任务往往是相关联的。比如行为类别之间存在着许多相关的信息。如图 1.2 所示，行为骑行和骑马的视觉特征非常相似，那么计算机可以借助一些骑马的特征去识别骑行。如果将这些被传统的单任务学习忽略的潜在信息加以利用，那么便可以建立更好的分类模型并提高模型的泛化能力。基于此，Rich Caruana 于 1997 年率先提出了多任务学习方法（Multi-task Learning）[25]。

多任务学习被提出后受到了广泛的关注和研究，目前它已经成功的运用到了很多的领域，例如：物体识别、语音识别、手写体识别和疾病进展预测等。在行为识别领域，多任务学习通过将每一个行为类别看做一个任务，通过利用行为类别之间潜藏的关系来提高识别的精度。下文会对多任务学习框架进行概述并详细介绍几种常用的多任务学习模型。

2.2　多任务学习方法概述

多任务学习方法建立在一个普遍的假设之上，即多个任务之间共享相同的特征

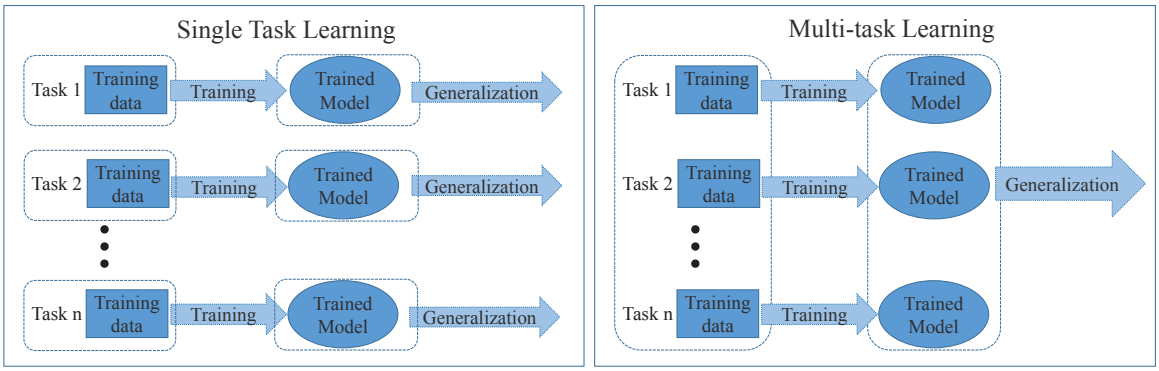


图 2.1　多任务学习与单任务学习的主要流程。

空间或者模型。在此假设上，多任务学习通过利用任务之间的潜藏信息来学习到更好且更具有泛化能力的模型。如图 2.1 所示，它与单任务学习的主要区别在于后者忽略了任务之间的关系，每次只学习一个模型。而多任务学习不仅仅考虑了任务之间的差异性，还考虑了任务的相关性。多任务学习特别适用于训练数据较少的情况，它通过利用数据之间的相关性来提高学习的性能。因此，多任务学习在计算机视觉领域得到了广泛的应用，它研究的重点在于如何寻找任务之间的关联性，以及如何将关联性与模型结合。

1997 年，Rich Caruana[25] 正式提出了多任务学习这一概念。他构造了一个前馈神经网络，这个网络中的每一个任务对应着一个输出节点，而所有任务都共享了隐藏节点，输入节点到隐藏节点权重参数的训练过程中利用了任务之间的潜在关联信息，而隐藏节点到输出节点的权重参数包含了每个任务特有的信息。尽管这个方法比较简单，但是却成功应用到了医疗诊断和车辆自动驾驶等领域。与共享参数类似的方法还有学习分层的贝叶斯模型参数[26]、高斯过程[27]的参数等。另一方面任务之间分享的不一定是参数，也可能是隐含的相同的特征空间。Evgeniou 和 Pontil[28] 假设所有的任务都是相关联的且都可以投射到一个共同的未知的空间中，并提出了一种多任务学习的正则化框架，通过基于 SVM 的方法将其转为传统的单任务学习求解。Andreas Argyriou[29] 将 $\ell_{2,1}$ 范数正则项与多任务学习结合，$\ell_{2,1}$ 范数正则项可以鼓励多个任务学习到相同的稀疏特征空间。他们首先对所有任务对应的参数建立参数矩阵，然后在 $\ell_{2,1}$ 范数的约束下最小化所有任务的经验风险值。

上述所有多任务学习方法都假设了数据库中的所有任务都是相关联的，然而这个假设对于现实的应用而言太过强烈，例如可能会存在一两个异常的类别，或如图 1.2 所示，类别之间的关系也是存在结构的。简单的假设所有任务共享所有的参数或共享一个特征空间可能会对模型的学习产生负面的影响。基于此，研究者们最近提出了许多任务学习方法来挖掘任务之间的潜在结构。Jacob 等人[30]考虑通过对任务进行自动聚类来确定它们之间的关联性，在同一个集群中的任务比不在同一个集群中任务的相关程度更高。在他们的工作中，任务之间的关联性被建模为参数的相似度。Zhuoliang Kang[31] 同样研究了任务的聚类问题，与 Jacob 工作的不同之处在于他们把任务之间的关联性建模为任务之间分享的特征。他们提出的算法可以同时学习任务之间潜在的结构和每个任务的分类参数。Pinghua Gong 等人[32]考虑到任务中异常点（outlier）的存在，于是提出了鲁棒的多任务特征学习（Robust Multi-Task Feature Learning, RMTFL），他们通过将分类参数矩阵分为两个组成部分，通过梯度下降法同时对相关联任务和异常任务学习分类器，并在理论上证明了该方法的可行性。Dinesh Jayaraman 等人[33]利用任务的先验信息对任务进行了分组，并将分组结构信息作为正则项与多任务学习框架结合来进行属性分类。上述的方法都假设了数

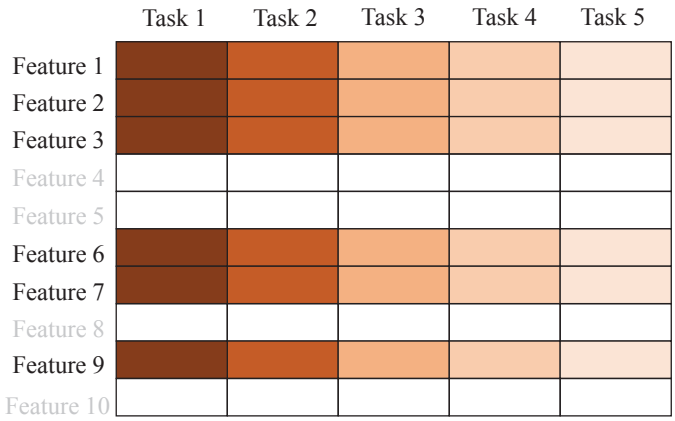


图 2.2　基于 $\ell_{2,1}$ 范数的多任务学习为所有的任务同时学习到一个低维的特征空间，无色的方块代表任务对于特征的权重为 0 。

据库中的类别可以聚类为不相交的组，每个组之里的任务可以投射到同一个维度的子空间，然而Abhishek Kumar 等人[34]认为上述假设对于真实世界的情况仍旧不太现实，因为位于不同组内的任务是不可能一点关系都没有的。所以他们通过学习一些隐任务来描述不同组内的任务之间共享的信息。

2.3　常用的多任务学习方法

2.3.1　基于 $\ell_{2,1}$ 范数的多任务学习

基于 $\ell_{2,1}$ 范数的多任务学习是非常经典和常用的一种多任务学习方法。它首先由Andreas Argyriou 等人[29]提出，之后被运用到了计算机视觉的很多领域，例如物体识别和人脸识别等。他们认为所有的任务都是相关的，且可以投影到一个被所有任务共享的低维特征空间。他们将可以为单任务产生稀疏性的 ℓ_1 范数扩展到了多任务的情况下。Andreas Argyriou 通过将该问题转化为等效的凸对偶形式从而优化求解。

基于 $\ell_{2,1}$ 范数的基本思想如图 2.2 所示，所有的任务都选择了相同维度的特征。假设数据库中有 C 个需要学习的任务，对于第 c 个任务有 N_c 个训练样本 $(\mathbf{x}_{c1}, y_{c1}), ..., (\mathbf{x}_{cN_c}, y_{cN_c}) \in \mathbb{R}^d \times \mathbb{R}$ ，该方法希望基于训练数据可以学习到 C 个分类或者回归函数 $f_c : \mathbb{R}^d \to \mathbb{R}$ 。函数 f_c 可以表示为：

$$f_c(x) = \sum_{i=1}^{d} w_{ic} x_i, \tag{2-1}$$

其中 w_{ic} 表示函数 f_c 对于第 i 维特征的分类或回归参数，x_i 表示样本 x 的第 i 维特征。基于 $\ell_{2,1}$ 范数的多任务学习算法通过让一些 w_{ic} 为 0 来达到选择特征的目的。则

目标函数为：

$$\min \sum_{i=1}^{N_c} L(y_{ci}, \mathbf{w}_c \mathbf{x}_{ci}) + \gamma \parallel \mathbf{w}_c \parallel_1^2, \tag{2-2}$$

其中 $\mathbf{w_c} \in \mathbf{R}^d$ 为 f_c 的参数向量，γ 为正则项参数。上式中用到的 ℓ_1 范数会为 $\mathbf{w_c}$ 产生稀疏的表示，将公式 (2-2) 扩展到多任务的情形下，可以得到目标函数为：

$$\min \sum_{c=1}^{C} \sum_{i=1}^{N_c} L(y_{ci}, \mathbf{w}_c \mathbf{x}_{ci}) + \gamma \parallel \mathbf{W} \parallel_{2,1}^2, \tag{2-3}$$

公式 (2-3) 是所有任务的损失函数，可以选择任意的一种具体损失函数来代替 L。$\mathbf{W} \in \mathbf{R}^{d \times C}$ 表示所有任务的参数矩阵，第二项是对分类矩阵 $\mathbf{W}$ 的约束。它对矩阵 $\mathbf{W}$ 的每一行计算 2 -norm，再对所得的结果计算 1 -norm。基于该方法的思想，学习到的 $\mathbf{W}$ 应该满足某一行全为 0 或不全为 0 ，如图 2.2 所示。全为 0 的行表示这一维特征对所有的任务都不重要，所以被抛弃。

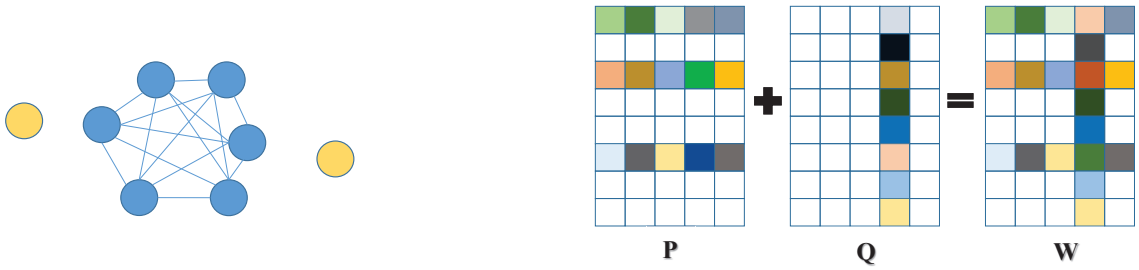
基于 $\ell_{2,1}$ 范数的多任务学习应用广泛，它特别适用于训练样本较少的情况，这时它可以充分的利用样本之间的信息提高性能和泛化能力。

2.3.2 基于组约束的多任务学习

基于 $\ell_{2,1}$ 范数的多任务学习可以为所有任务学习到一个低维的特征空间，但是由于它建立在所有任务都相关的假设之上，导致了它不适用于绝大多数的现实问题。基于此，研究者们提出了一些基于组约束的多任务学习方法。即任务可以根据相似程度分为若干个组，位于同一个组内的任务分享相同的特征空间。下面将介绍几种典型和常用的基于组约束的多任务学习方法。

（ 1）Pinghua Gong[32] 于 2012 年提出了鲁棒的多任务特征学习（RMTFL)，他认为数据库中大部分任务都是彼此相关联的，而那些不相关的任务称为异常任务，如图 2.3(a) 所示。如果按照基于 $\ell_{2,1}$ 范数的多任务学习忽略异常任务的存在，那么异常任务反而会降低模型的性能。RMTFL 可以同时分别为相关联的任务和异常任务学习到各自的特征空间。并且此方法不需要预先知道哪个任务是异常的，而是根据数据的特征自动的将它学习出来。RMTFL 将所有任务的参数权重矩阵 $\mathbf{W}$ 分解成了 $\mathbf{P}$ 和 $\mathbf{Q}$ 两个元素之和，对 $\mathbf{P}$ 加以 $\ell_{2,1}$ 范数的约束使得相关的任务学习到相同的低维特征空间，同时也对元素 $\mathbf{Q}$ 的转置也加以 $\ell_{2,1}$ 约束从而学习到异常任务对应的模型参数，RMTFL 具体的学习框架如下所示：

$$\begin{aligned} \min \sum_{c=1}^{C} \sum_{i=1}^{N_c} L(y_{ci}, \mathbf{w}_c \mathbf{x}_{ci}) + \lambda_1 \parallel \mathbf{P} \parallel_{2,1}^2 + \lambda_2 \parallel \mathbf{Q}^T \parallel_{2,1}^1, \\ s.t.\ \mathbf{W} = \mathbf{P} + \mathbf{Q}, \end{aligned} \tag{2-4}$$



(a)　　(b)

图 2.3　(a)在大部分真实情况下，不一定是所有任务都彼此关联，总有少数异常任务的存在，他们不与其他任务共享特征空间。(b)RMTFL 的矩阵分解示意图。方块代表了某一个任务对于特征的权重，白色为 0，彩色为非 0。

如公式 (2-4) 所示，第二个关于 $\mathbf{P}$ 的正则项为相关任务学习到相同的特征空间，第三个关于 $\mathbf{Q}$ 的正则项可以发现异常任务。参数 λ_1 和 λ_2 分别控制两个正则项的惩罚力度。事实上通过观察可以发现，公式 (2-4) 的前两项其实就是公式 (2-3) 的变形，正如 2.3.1 小节所述，对元素 $\mathbf{P}$ 加以 $\ell_{2,1}$ 的约束可以使 $\mathbf{P}$ 的行包含全 0 或者全不为 0 的元素，从而为所有相关的任务学习到相同的特征空间。然而由于异常任务的存在，假设所有的任务都共享同一个特征空间不太现实，所以又引入了 (2-4) 中的第三项。通过对元素 $\mathbf{Q}$ 的转置加以 $\ell_{2,1}$ 的约束可以使 $\mathbf{Q}$ 的列为全 0 或全不为 0，而全部不为 0 的那一列就对应着异常任务对于所有特征维度的权重系数。如图 2.3(b) 所示，如果 $\mathbf{Q}$ 的第 i 列全不为 0，那么对应的 $\mathbf{W}$ 的矩阵也应当全不为 0，这样第 i 个任务作为一个异常任务不与其他任务共享特征空间，同时 $\mathbf{Q}$ 中其他全为 0 的列代表了这些任务彼此相关，而它们的模型参数可以通过求解 $\mathbf{P}$ 得到。

（2）Abhishek Kumar[34] 提出了一种新的基于组约束的多任务学习算法。众所周知多任务学习建立在任务之间分享信息的基础之上，那么可以直接把任务之间共享的信息学习出来，Abhishek Kumar 把任务之间共享的信息称之为基任务，并假设每一个任务可以表示为基任务的线性组合。这个算法与 RMTFL 等算法的优越性在于，RMTFL 认为相关任务群和异常任务不分享任何信息，然而对于现实情况而言任务与任务之间完全没有关系是很少见的，这个算法通过学习到有限的基任务再将任务表示为基任务的组合，可以使不同任务之间按亲疏程度共享一定数目的信息。如果两个任务关系密切，那么选中的相同的基任务的数量也会比较多。该算法将所有任务对应的参数矩阵 $\mathbf{W}$ 表示为 $\mathbf{L}$ 和 $\mathbf{S}$ 两个元素之积，$\mathbf{L} \in \mathbb{R}^{d\times k}$ 的每一列都表示基任务，$\mathbf{S} \in \mathbb{R}^{k\times C}$ 表示了 C 个组合分别对于基任务的权重，具体框架如下所示：

$$\begin{aligned} \min \sum_{c=1}^{C}\sum_{i=1}^{N_c} L(y_{ci}, \mathbf{w}_c\mathbf{x}_{ci}) + \lambda_1 \parallel \mathbf{S} \parallel_1 + \lambda_2 \parallel \mathbf{L}^T \parallel_F^2, \\ s.t.\ \mathbf{W} = \mathbf{LS}, \end{aligned} \tag{2-5}$$

如上式所示，第二项通过对 $\mathbf{S}$ 加以 ℓ_1 的约束，可以使 $\mathbf{S}$ 矩阵的每一列产生稀疏表

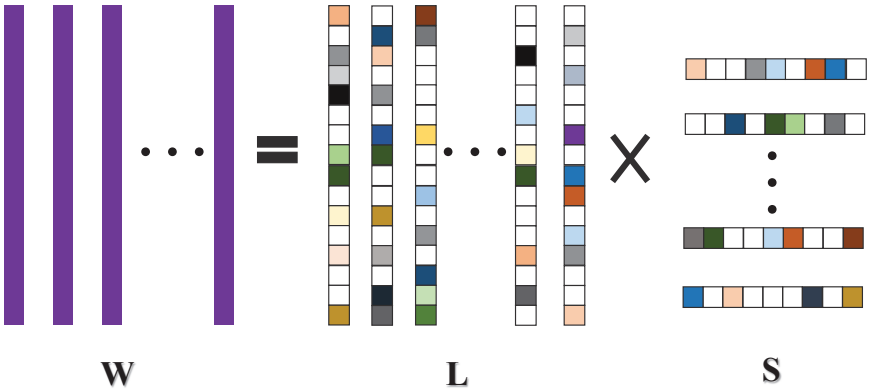


图 2.4 将权重矩阵 **W** 表示为矩阵 **L** 和矩阵 **S** 的积。 **L** 表示了任务之间分享的基任务， **S** 包含了数据库中任务潜在的结构信息。

示，从而鼓励每一个任务可以仅被最重要和包含最多信息的基任务表示。$\| \mathbf{L}^T \|_F^2$ 可以防止矩阵 **L** 过拟合，它的每一列都代表了被 d 维特征所描述的基任务。矩阵 **S** 中的任意两行 $\mathbf{s}_{c1}$ 与 $\mathbf{s}_{c2}$ 对应着任务 $c1$ 和任务 $c2$ 选中的基任务。$\mathbf{s}_{c1}$ 与 $\mathbf{s}_{c2}$ 的相似程度代表了这两个任务的相似程度。如果 $\mathbf{s}_{c1}$ 与 $\mathbf{s}_{c2}$ 具有完全相同的稀疏形式，说明任务 $c1$ 和任务 $c2$ 属于同一组。如果 $\mathbf{s}_{c1}$ 与 $\mathbf{s}_{c2}$ 的稀疏形式部分重叠，说明任务 $c1$ 和任务 $c2$ 属于不同的组，但是允许关联性不是很大的任务之间也分享特征信息。如果一个任务不与其他任何任务共享基任务，那么这个任务就是RMTFL的异常任务，如图 **??** 所示, **S** 中也蕴藏了数据库中任务潜在的结构信息。

公式 (2-5) 在 **S** 和 **L** 分别固定的情况下是凸函数，所以可以使用交替迭代的算法来求解 **S** 和 **L** 的局部最小值。

Qiang Zhou[35] 等人在上述方法的基础上做出了改进，并将此多任务学习方法用于了行为识别。这里任务之间分享的基任务可以视为不同的行为类别之间共享的运动元素。比如跑步、走路、挥手等运动其实都是有身体部位的基本移动构成的。最终每个类别的分类器参数由基本运动元素线性组合而成。具体框架如下所示：

$$\begin{aligned} \min \sum_{c=1}^{C} \sum_{i=1}^{N_c} L(y_{ci}, \mathbf{w}_c \mathbf{x}_{ci}) + \lambda_1 \| \mathbf{S} \|_1 + \lambda_2 \| \mathbf{L}^T \|_F^2 + \lambda_3 \| \mathbf{L} \|_1, \\ s.t.\ \mathbf{W} = \mathbf{LS}, \end{aligned} \tag{2-6}$$

与公式 (2-5) 的差别在于，公式 (2-6) 引进了对于基任务 **L** 的 ℓ_1 约束。因为在行为识别中，基任务应该对应的是被行为类别之间共享最多的运动元素，如果全局信息太多的话反而会降低判别性。对 **L** 矩阵加以 ℓ_1 的约束使得 **L** 的行具有稀疏表示，而不为 0 的项对应着最重要的基本运动元素，这也是一种特征选择。

Qiang Zhou[35] 运用加速近似梯度下降算法的求得公式 (2-6) 后，实验结果表明了与单任务学习相比多任务学习可以有效的提高行为识别的精度。这说明了行为类别

之间分享着许多的信息，这也从侧面证明了本论文的工作的必要性。

（3）上述讨论的任务之间的结构都是单层的，Seyoung Kim[36] 等人认为任务之间可能存在多层的结构，比如图 2.5(a) 所示的树结构或者图结构。基于此，他们提出了树形结构正则项和基于树形结构约束的多任务学习。与上文中介绍的方法相似，基于树形结构约束的多任务学习也允许任务在多重的粒度上分组，即允许关联性不是很大的任务之间也分享特征信息，但是基于树形结构约束的多任务学习不会出现由于分享信息的重叠导致的惩罚过度的问题。如图 2.5(a) 所示，假设任务之间的关系可以表示为拥有 V 个节点的树形结构。叶子节点代表任务，中间节点代表若干个任务构成的组，同属于一个节点的任务之间的相似度更高且被鼓励共享特征信息，据树底越近的中间节点表示它拥有的子树的关系越密切。为了量化每个组的关联程度，Seyoung Kim 给每一个节点 v 都赋以权重 s_v，高度越低的节点的 s_v 越小。将此树形结构用数学的形式表达即可得到树形结构正则项：

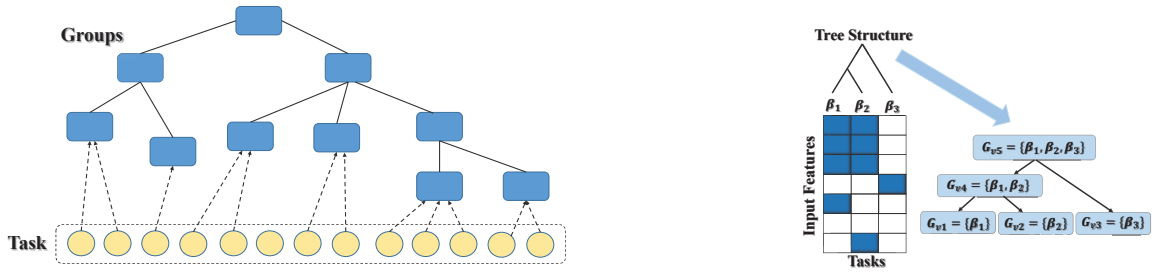
$$\sum_{j}\sum_{v\in V} s_v \parallel \mathbf{w}^j_{G_v} \parallel_2, \tag{2-7}$$

其中 $\mathbf{w}^j_{G_v} = \mathbf{w}^j_k : k \in G_v$ 表示属于第 v 组中所有任务对于第 j 维特征参数向量。而 s_v 保证了不会有参数被过度惩罚。在实际的运算中组的结构和每个节点对应的权重可以通过层次聚类算法得到。将树形结构正则项与多任务学习结合可以使任务选择的特征结构也呈对应的树形结构。如图 2.5(b) 所示，假设输入的数据库中有 3 个任务，且通过树形聚类可得到 4 个分组，即 $G_{v1} = \mathbf{w}_1, G_{v2} = \mathbf{w}_2, G_{v3} = \mathbf{w}_3, G_{v4} = \{\mathbf{w}_1, \mathbf{w}_2\}$，其中 $\mathbf{w}$ 为各个任务对应的分类参数，学习到的模型参数矩阵也对应着树形结构，即首先任务 1 和任务 2 分享特征，而任务 1、任务 2、任务 3 之间又互相竞争和相斥，基于树形结构约束的多任务学习具体框架如下所示：

$$\min \sum_{c=1}^{C}\sum_{i=1}^{N_c} L(y_{ci}, \mathbf{w}_c x_{ci}) + \lambda \sum_{j}\sum_{v\in V} s_v \parallel \mathbf{w}^j_{G_v} \parallel_2, \tag{2-8}$$

与上文中介绍的方法相比，基于树形结构约束的多任务学习在对于任务分组和任务信息分享上更加的灵活，它不仅仅允许属于不同组的任务共享信息，还通过对每个组加入权重使每个任务都得到均衡的惩罚，然而它的缺点在于任务之间的组的结构信息属于先验知识，即需要先用一种距离度量出任务之间的相似性，再用聚类方法得到树形结构，而其他方法可以同时优化出任务的结构和最终的模型的参数矩阵。

（4）尽管基于树形结构约束的多任务学习在任务分组和信息共享十分灵活，但是大部分情况下任务的结构还是单层的。基于此，Dinesh Jayaraman[33] 对基于树形结构约束的多任务学习做出了改进，提出了一种既可以灵活共享特征又是单层的组结构的多任务学习方法，并将其用于属性学习，在属性学习中将每个属性视为任



(a) (b)

图 2.5 (a)任务之间存在树形结构，圆圈表示任务，方块表示组。同属于同一个组的任务相似度比较高。(b)任务之间存在树形结构，圆圈表示任务，方块表示组。同属于同一个组的任务相似度比较高。

务，且属性之间具有显而易见的组结构。例如描述颜色的属性同属一组，它们之间共享描述颜色的特征。这个目标通过对任务组内加以 $\ell_{2,1}$ 约束和组间加以 ℓ_1 约束的方式实现，该方法的具体的目标函数如下：

$$\min \sum_{c=1}^{C} \sum_{i=1}^{N_c} L(y_{ti}, \mathbf{w}_c \mathbf{x}_{ci}) + \lambda \sum_{d=1}^{D} \sum_{v \in V} \| \mathbf{w}_{G_v}^d \|_2, \tag{2-9}$$

这里 $\mathbf{w}_{G_v}^d$ 是一个行向量，它代表了属于第 v 组中所有任务对于第 d 维特征的参数向量。第二项正则项对 $\mathbf{W}$ 矩阵逐行惩罚，先在每个任务组内计算 $\ell_{2,1}$ 范数，再在组间计算 ℓ_1 范数，从而鼓励同属于一组内的任务选择相同的特征维数，达到组内共享与组间竞争的期望。如果数据库中的 C 个任务都属于同一个组，即数据库中只有一个任务组，此时 $G_1 = 1, 2, ..., C$ ，组结构约束也退化为 $\ell_{2,1}$ 范数，公式 (2-9) 相应的也退化为 2.3.1 小节中的基于 $\ell_{2,1}$ 范数的多任务学习框架，鼓励所有任务学习到一个特征空间。如果数据中的的 C 个任务互不相关，那么就有 C 个任务组 $G_1, G_2, ...G_C$ ，每个组内只包含一个任务。此时组结构约束退化为 ℓ_1 范数，公式 (2-9) 将为所有的任务学习出各不相同的稀疏的特征表示。

2.4 本章小结

单任务学习为每个任务学习独立的模型，忽略了任务之间的关系。多任务学习通过利用任务之间共享的信息弥补了单任务学习的这一缺陷，并且提高了模型的性能和泛化能力。目前为止研究者们提出了很多的多任务学习方法框架，主要分为基于 $\ell_{2,1}$ 范数的多任务学习和基于组约束的多任务学习。前者也可以称为联合学习的多任务学习方法，因为它为所有的任务联合学习同一个低维度的特征空间并鼓励所有任务之间都分享特征信息。后者认为基于 $\ell_{2,1}$ 范数的多任务学习方法不适用于绝大多数的现实情况。于是研究者们提出了各种方法来探索任务之间的结构信息并用

它们来约束多任务学习的过程。其中典型的有RMTFL等单层结构模型，还有树形结构和图模型结构等多层模型。这一类方法将数据库中的任务分为若干组，相似的任务位于同一组。同一组内的任务存在特征共享，而不同组内的任务存在特征竞争。

第三章　基于SVM间隔的多任务行为识别

3.1　引言

随着计算机网络不断发展和 YouTube 、秒拍等视频社交网站和 APP 的飞速普及，视频数据已逐渐成为人们获取信息的主要载体。视频数目飞快增长的同时，其包含的内容也愈加丰富。视频与其他媒体介质相比，首先获取成本相似，且在信息传播的深度和广度上都更有优势。在这种背景之下，对视频中的行为进行识别具有极大的理论研究价值及广阔的应用市场。如第一章所述，行为识别实际上是一个分类问题，现有的行为识别方法的一般流程都是先从视频中提取可以表示人体行为的特征，然后利用特征训练分类器。然而目前的方法把每一个行为类别当做了一个任务，对每个任务学习模型的过程是独立的，从而忽略了任务之间的信息。如何利用行为类别之间的信息，提高行为识别的性能是本文的研究重点。

本论文通过观察发现数据库中的行为类别存在潜在的结构信息，如图 1.2 所示，在 HMDB51 数据库中的行为咀嚼（chew）和吃饭（eat）非常相似，所以它们同属于一组且分享特征，反之咀嚼和骑马（ride horse）不属于同一组所以不共享特征信息。本论文基于这个观察并针对现存方法不能有效利用行为类别之间信息的不足，提出了利用基于组约束的多任务学习框架来实现行为的识别。具体的流程框架如图 3.1 所示，首先对数据库中的所有任务两两训练 1-VS-1 的 SVM ，如果数据库中共有 n 个种类，即总共训练 C_n^2 个 SVMs ，然后用 SVM 的间隔来度量对应的两个类别之间的相似性程度，对相似度矩阵进行聚类可以得到数据库的潜在的类别结构，最后将此结构作为先验约束，建立目标函数，求解优化得到所有任务对应的分类参数，完成行为的识别。

本章提出了一种基于 SVM 间隔的多任务行为识别方法，主要内容安排如下：在 3.2 节中介绍任务之间的相似性距离和任务潜在结构的计算方法；在 3.3 节中介绍目标函数和优化；在 3.4 节中进行实验仿真并对实验结果进行分析；在 3.5 节中对本章内容进行总结。

3.2　基于SVM间隔的多任务行为识别

3.2.1　SVM间隔

支持向量机（Support Vector Machines, SVM）[37][38]是一种二分类模型，它可以在合适的特征空间中为训练样本寻找到间隔最大的分类超平面。Cortes 与 Vapnik[37]

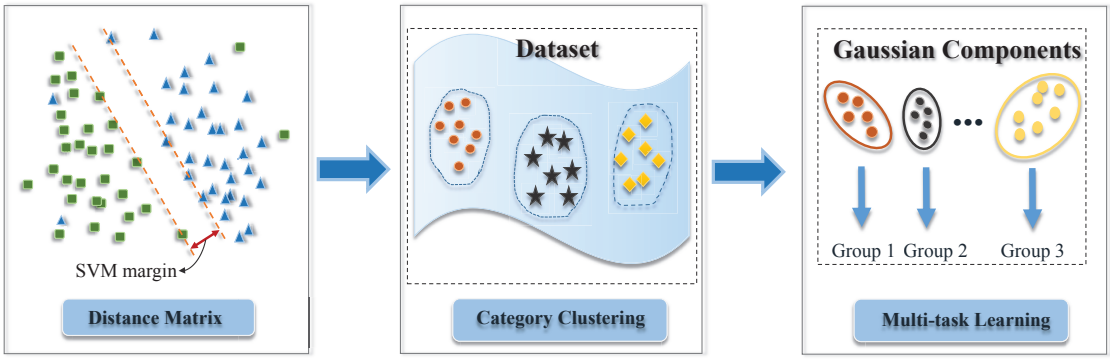


图 3.1　基于 SVM 间隔的多任务行为识别的主要流程。

于 1995 年首先提出了线性支持向量机的概念，随着核函数（kernel function）的引入，学者们提出了非线性的支持向量机。当输入空间为欧式空间与离散集合、特征空间为希尔伯特空间时，核函数表示将输入从输入空间映射到特征空间得到的特征向量之间的内积。一个线性的支持向量机的目的是从n 维的特征空间中找到一个超平面，如图 3.2 所示，该超平面可以表示为：

$$w^T x + b = 0. \tag{3-1}$$

一个超平面在二维空间中就是一条直线，SVM 希望通过这条直线将两类数据分割开来。具体来说 SVM 希望根据训练数据找到合适的 w 和 b，令 $f(x) = w^T x + b$，要求所有在直线 $w^T x + b = 0$ 一侧的样本对应的标签为 1，另一端的样本标签为 -1。如图 3.2 所示，两种颜色的点分别表示两个种类，红色直线表示法向量为 w 的超平面，每个样本点据超平面的距离可以表示它分类正确的确信度。即越接近超平面的点越不容易被分类，反之，据超平面很远的点则很容易分辨出其类别。从图中可以观察到，可以将这两种数据线性分类的超平面很多，即可行的 w 和 b 不是唯一的。正如上文提到的，SVM 中定义了最好的超平面是分类间隔最大的超平面。假设已给定训练集 X 和超平面 $w^T x + b = 0$，则定义超平面 (w, b) 关于样本点 (x_i, y_i) 的函数间隔为：

$$\hat{\gamma}_i = y_i(w^T x_i + b), \tag{3-2}$$

超平面 $w^T x + b = 0$ 对于给定训练集 X 的函数间隔被定义为超平面对于 X 中所有样本 (x_i, y_i) 的函数间隔的最小值，即：

$$\hat{\gamma} = \min_{i=1,\ldots,N} \hat{\gamma}_i. \tag{3-3}$$

函数间隔在一定程度上可以表示分类的精确度和确信度。但是在选择超平面时，仅仅有函数间隔还是不够的。因为如果成比例的改变 w 和 b 为 $3w$ 和 $3b$，函数间隔

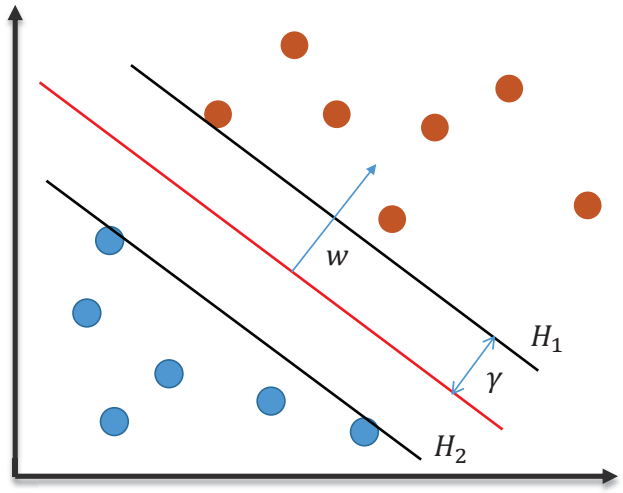


图 3.2　最大几何间隔分类超平面示例，据超平面（红色线）最近的样本点称为支持向量。

变成了原来的 3 倍，但是事实是超平面却没有改变。为了避免这一事实，可以通过对超平面的法向量 w 加一些规范化的约束，例如限制 $\| w \| = 1$，使得无论 w 和 b 无论怎么变化，间隔都是不变的。这是函数间隔变成了几何间隔。具体来说超平面 $w^T x + b = 0$ 关于样本点 (x_i, y_i) 的几何间隔为

$$\hat{\gamma}_i = y_i(\frac{w^T}{\| w \|} x_i + \frac{b}{\| w \|}), \tag{3-4}$$

和函数间隔的定义类似，超平面 $w^T x + b = 0$ 对于给定数据集 X 的几何间隔被定义为超平面对于 X 中所有样本 (x_i, y_i) 的几何间隔之最小值，即

$$\gamma = \min_{i=1,...,N} \gamma_i. \tag{3-5}$$

SVM 的目的即根据训练样本寻找即能正确划分数据集且几何间隔最大的超平面。对于一个训练数据集而言，可以线性分割数据的超平面可能有很多个，但是几何间隔最大的超平面只有一个，它与其他的超平面相比，它的泛化能力更强，对于未知标签的测试用例有较好的分类预测能力。SVM 的具体的损失函数如下所示：

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \| w \|^2 \\ &s.t.\ y_i(w^T x_i + b - 1) \geq 0, i = 1, 2, ..., N \end{aligned} \tag{3-6}$$

如图 3.2 所示，支持向量被定义为距离超平面最近的样本点。对于 $y_i = 1$ 的样本点，支持向量在超平面 $H_1 : w^T x + b = 1$ 上，对于 $y_i = -1$ 的样本点，支持向量在超平面 $H_2 : w^T x + b = -1$ 上。在 H_1 和 H_2 上的点就是支持向量。用数学的语言来讲的话，支持向量是令公式 (3-6) 成立的样本点，即满足

$$y_i(w^T x_i + b - 1) = 0. \tag{3-7}$$

注意到 H_1 和 H_2 平行，并且没有样本点落在这两条直线之间，H_1 和 H_2 之间的距离称为间隔。间隔的大小依赖于超平面的法向量 w，等于 $\frac{2}{\|w\|}$。H_1 和 H_2 称为间隔边界。

在决定超平面时只有作为支持向量的样本点才起作用，而其他样本点并没有帮助。如果移动支持向量将改变超平面的位置，但是如果在间隔边界以外移动其他任意样本点，甚至去掉或者添加新的样本点，也对最优超平面不起任何影响。支持向量机名字的由来也是因为支持向量在确定超平面时起着决定性的作用。如果两个类别的差异性很大，很好被区分，那么由训练样本训练得到的 SVM 的间隔必定也很大，反之如果两个类别非常相似，那么得到的 SVM 的间隔也会很小。所以可以用 SVM 的间隔来衡量两个类别之间的相似度。

3.2.2 基于 SVM 的相似性度量

由上文所述，无论是真实世界还是数据库中，不同的行为类别之间都是有关系的。进一步通过观察发现行为识别的数据库中的行为类别存在潜在的结构信息，具体来说我们认为数据库中行为可以分为几个组，相似的行为类别同属于一组且分享特征，反之不属于于同一组的类别所以不共享特征信息。在上一小节的基础上，我们选择了 SVM 间隔作为类别之间的相似性度量，通过计算所有类别之间的相似度矩阵聚类之后得到类别之间潜在的数据结构。

为了将数据库中任意两个行为类别的相似程度量化，我们利用相对应的视频数据对每对行为类别训练 1-VS-1 的SVM。正如上一个小节中提到的，两种行为如果越相似，那么它们对应的 SVM 间隔就会越小，说明它们越难被线性分类。我们设 s_{c_p,c_q} 为类别 p 和 q 之间的相似度，则 s_{c_p,c_q} 的具体计算公式如下:

$$s_{c_p,c_q} = \sum_{\forall(i,j)\in c_p\cup c_q} \alpha_i\alpha_j(-1)^{I[c(i),c(j)]}K(x_{p,i},x_{q,i}), \tag{3-8}$$

上式中，i 和 j 表示来自第 p 个类别和第 q 个类别交集的两个样本序号。$x_{p,i}$ 表示来自第 p 个类别的第 i 个样本。α 表示支持向量的系数。α 可以通过训练 SVM 时通过求解最优超平面得到，每一个训练样本都有一个对应的 α 值，只是仅仅当这个训练样本是支持向量时 α 的值不为 0。$c(.)$ 可以返回给定样本的类标序号。$I[.]$ 也是一个函数，它的具体形式如下：

$$I[c(i),c(j)] = \begin{cases} 0 & otherwise \\ 1 & c[i]\neq c[j], \end{cases} \tag{3-9}$$

当样本 i 和 j 具有相同的类标时，$I[c(i),c(j)]$ 的值为 0，因为我们要计算的是两个类别之间的相似度，当 i 和 j 来自同一个类的时候是没有意义的。

$K(.,.)$ 是核函数。核函数在线性模型不可分的情况下可以通过一个非线性变换将训练样本从输入空间映射到一个线性可分的高维空间，从而在高维的空间中寻找几何间隔的超平面。核函数有很多的种类，常用的有多项式核函数（polynomial kernel function）、高斯核函数（Gaussian kernel function）和线性核函数等。在公式 (3-8) 中，核函数的选择应该视特征的种类而定。在本文中我们使用的特征为Fisher Vector，所以核函数为线性核函数。而且由于每个 1-VS-1 的 SVM 只需要用到两个类别对应的训练样本，所以即使总共需要训练 $\binom{2}{C}$ 个分类器，但是速度还是很快的。

通过对数据库中的每对类别训练 1-VS-1 的 SVM 并计算根据公式 (3-8)，最终可以得到数据库中所有类别对应的相似度矩阵 $\mathbf{S} \in \mathbb{R}^{(C\times C)}$，这是一个对称矩阵，且 $s_{p,q}$ 代表了第 p 个类别和第 q 个类别的相似度，$s_{p,q}$ 越小，说明这两个类越相似，反之亦然。得到相似度矩阵 $\mathbf{S}$ 之后，我们利用近邻传播聚类（Affinity Propagation Clustering）[39]得到了数据库中隐含的类别的结构信息。值得注意的是，我们不需要提前决定聚类结构中簇的数量，它可以基于相似度矩阵 $\mathbf{S}$ 被聚类算法所决定。这么做的原因有两点，首先凭肉眼或者人工的感觉来判断行为类别相似是不准确的，视觉信息不相似的行为类别有时也会共享相似的运动轨迹等特征。所以人工的判断数据库中行为组的个数是不准确的，不仅如此，工作量还非常的大。此外，根据公式 (3-8) 可以看出类别之间相似性的度量是基于特征计算的。具体来说，对于不同的特征，例如方向梯度直方图（Histograms of Oriented Gradients, HOG）和光流场方向直方图（Histograms of Optical Flow, HOF），得到的 SVMs 和相似度矩阵 $\mathbf{S}$ 也是不同的，那么随之对应的类别的组结构也是不同的。总而言之，对于一个行为识别数据库并不存在一个最优的组结构，组的个数和每个组内具体的行为类别是随特征类型的变化而变化的。

在 Hou Rui[40] 的工作中，也利用到了基于 SVM 间隔的相似性度量。但是与他们工作相比，我们的方法在应用和构造上都有所创新改变。Hou Rui 通过计算 SVM 间隔来找到每个类别的相似类别，从而可以学习到高级语义上的判别性特征。然而我们通过计算基于 SVM 间隔的相似度来探索潜藏在数据库中的类别结构信息，从而可以用其约束多任务学习鼓励同一个组内的类别共享特征信息。

3.2.3　目标函数的建立和优化

基于SVM间隔的多任务行为识别的具体流程如图 3.1 所示。如果将数据库中的行为类别视为任务，通过观察我们发现这些任务之间是有关系的。而且我们假设任务存在潜在的结构信息，即任务可以分为若干个组，每个组内的任务都是相似的且共享相同的特征空间，不同组内的任务进行特征排斥和竞争。为了求得任务的结构信息，我们通过对每对任务训练 1-VS-1 的 SVM，利用公式 (3-8) 求得每对任务之间

的相似性距离，得到相似性矩阵 $\mathbf{S}$，然后通过聚类得到基于训练特征的最优的任务结构。我们把该结构作为先验信息，利用基于组结构约束的多任务学习框架同时学习每个任务对应的二元线性分类器。

假设在给定行为数据库中共有 C 类行为类别。对于第 c 类别，对应的训练样本为 $\{x_{c,i}, y_{c,i}\}_{i=1}^{N_c} \subset \mathrm{R}^{D\times 1}(c = 1, ..., C)$，其中 i 表示第 c 个类别的视频样本的序号，N_c 表示属于第 c 个类别的样本总量。$x_{p,i}$ 表示来自第 p 个类别的第 i 个样本的 D 维特征，$y_{c,i}$ 为其对应的标签。且假设通过聚类之后可将任务总共分为 L 组。则具体的目标函数如下：

$$\min_{W} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \frac{1}{2} \left[\max(0, 1 - y_{c,i} w_c x_{c,i})\right]^2 + \lambda_1 \sum_{d=1}^{D} \sum_{l=1}^{L} \|w^{d,g_l}\|_2 + \lambda_2 \|W\|_F^2, \tag{3-10}$$

上式中的第一项为铰链损失函数（hinge loss function）。$W \in \mathrm{R}^{N\times D}$表示所有任务对应的二元线性分类器的参数矩阵，$w_c \in \mathrm{R}^D$表示第$c$ 个任务的参数。第二项为基于组结构的正则项，其中w^{d,g_l}是一个行向量，它表示属于第g_l个组的所有任务对应第d维特征的权重参数。基于组结构的正则项同时利用了ℓ_1范数和$\ell_{2,1}$范数的优点，从而鼓励组内的特征共享和组间的特征竞争。最后一项为F范数（Frobenius norm）可以防止分类矩阵W过拟合。λ_1和λ_2是用来控制稀疏与分类损失函数之间的平衡因子。

很明显可以看出，用以约束公式(**??**)的组结构正则项$\sum_{d=1}^{D} \sum_{l=1}^{L} \|w^{d,g_l}\|_2$是一个混合范数正则项。尽管它是凸的，但是它优化起来却非平稳

优化求解公式(**??**)之后可以得到每一个行为类别对应的二元线性分类参数w_c。这时对于一个新的测试视频，首先先提取对它提取特征，将它表示为$x_n \in \mathbf{R}^D$，它对应的标签y_n可由以下的公式得到：

$$y_n = \arg\max_{c} w_c x_n .? \tag{3-11}$$

Dinesh Jayaraman[33]等也利用了基于组结构约束的多任务行为，他通过将属性视为任务来同时对每个属性学习分类模型。基于SVM间隔的多任务行为识别方法与他们的工作对比，有如下不同之处：（1）首先前者中任务的组结构是人工分割的。Dinesh Jayaraman通过将描述同类事物的属性分为一组，例如例如描述颜色的属性同属一组，它们之间共享描述颜色的特征。这种做法先验假设太强，而且可能忽视任务之间潜藏的肉眼不可见的信息，所以我们的方法提出了利用任务之间的SVM间隔来衡量任务之间的相似度。（2）我们的方法利用了ℓ_2范数来防止分类矩阵过拟合，提高了模型的泛化能力。

图 3.3　HMDB51中的样本示例。

3.3　实验结果与分析

本小节先介绍了实验中所用数据库、特征、基准方法和参数选择等细节。然后展示了基于SVM间隔的多任务行为识别和对比方法分别在行为识别数据库HMDB51的结果，并给出了结果的分析。

3.3.1　实验设计

为了证明基于SVM间隔的多任务行为识别方法是有效的，我们将在行为识别数据库HMDB51[41]上与多种基准方法与目前行为识别领域最先进的一些方法做对比，下面将依次介绍数据库、特征、参数和基准方法。

HMDB51数据库中总共包含6766个视频，它们分别来自51个行为类别，且平均每个类别包含100个以上的视频样本。如图所示所有的样本都是在真实环境下拍摄而得的，比如电影、视频网站等。由于视角、尺寸、背景和光照等因素，HMDB51中即使属于同一类的视频差别也非常大。HMDB51对于行为识别应用而言，是个非常困难的数据库。在数据库的主页上提供了三种训练集和测试集的分割方法，我们的实验也建立在这三种数据分割方法之上，并且最终给出了平均分类精度。

在我们的实验中，我们利用了improved dense trajectories[16]来提取局部描述子。首先在连续帧中跟踪兴趣点，然后在轨迹的附近的范围提取HOG、HOF、MBHx和MBHy 四种类型的描述子。和[42]中的步骤一样，我们首先利用PCA将每种描述子的维度降到一半。然后随机选取256000个特征，通过EM算法学习到由512个高斯组成的高斯混合模型，从而将描述子转化为Fisher Vector的形式。然后对Fisher Vector进行白化[43]、L_2正规化和intra正规化之后，得到了最终的特征。在我们的实验中，分别给出了四种类型的描述子单独训练的情况，也给出了将这四种特征串联起来的情况。

在公式(**??**)中的λ_1和λ_2的最优值通过交叉验证选出。而数据库的类别组的最优个数由反向传播算子根据不同的特征得到。

特征/方法	多任务行为识别	基于ℓ_1范数的多任务行为识别	基于ℓ_{21} 范数的多任务行为识别	基于SVM间隔多任务行为识别
HOG	40.22	44.05	43.84	**44.54**
HOF	49.00	50.04	50.09	**50.62**
MBHx	40.24	42.81	42.61	**44.09**
MBHy	47.13	48.60	48.62	**49.01**
combined	60.15	60.38	60.31	**60.54**

表 3.1 基于SVM间隔的多任务学习与基准方法在HMDB51上的实验结果对比。

我们提出了3种基准方法，分别为不加正则项的多任务行为识别，基于ℓ_1范数的多任务行为识别，基于ℓ_{21}范数的多任务行为识别。在所有的多任务行为识别方法中，我们选择Hinge loss作为损失函数。

3.3.2 实验结果

首先我们得到了当四种特征串联时，通反向传播算子聚类方法将51个行为类别划分为了7个组，具体结构为组1: *turn, walk, shake hands, hug and kiss.* 组2: *brush hair, clap, wave, shoot gun, draw sword, sword, climb, climb stairs, drink, eat, pick, sit, stand, pour, shoot bow and pullup.* 组3: *chew, smile, laugh, smoke and talk.* 组4: *push up and sit up.* 组5: *cartwheel, flic flac, hand stand, somersault, push, ride bike and ride horse.* 组6: *catch, hit, swing baseball, throw, dive, fall floor, jump, run, kick ball, fencing, kick, sword exercise and punch.* 组7: *dribble, shoot ball and golf.* 通过这个结果可以看出在视觉和行为轨迹上相似的行为都被分到了一组。比如，行为chew, smile, laugh, smoke and talk这些描述面部的微小特征的行为被分到了同一组。

然后我们比较了基于SVM间隔的多任务行为识别和基准方法在HMDB51上的识别精度，具体结果如表3.1所示。通过结果可以看出基于SVM间隔的多任务行为识别无论用哪种类型的特征，效果都比其它的多任务学习框架要好。当所有的行为类别都属于同一个组的时候，也就是所有的类别都相似的情况下，公式(**??**)中的结构正则项退化为$\ell_{2,1}$范数。当所有的行为类别都自成一组时，也就是所有的类别之间都不共享特征的情况下，结构正则项退化为ℓ_1范数。所以组结构约束综合了ℓ_1范数和$\ell_{2,1}$范数的优点。

最后我们把基于SVM间隔的多任务行为识别和行为识别领域一些先进的方法在HMDB51数据库上做了对比试验。如表3.2所示，基于SVM间隔的多任务行为识别甚至比基于深度学习的方法[44]效果还要好。

Methods	Accuracy
Sadanand and J.Corso[45]	26.9%
Yang et al.[46]	53.9%
Hou et al.[40]	57.88%
K. Simonyan and A. Zisserman[44]	59.5%
基于SVM间隔多任务行为识别	**60.54**%

表 3.2　基于SVM间隔的多任务行为识别方法与行为识别领域一些先进方法在HMDB51上的实验结果对比。

3.4　本章小结

之前绝大部分的人体行为识别方法都将行为类别当做了独立的任务，分别为每个类别训练其对应分类器。然而通过观察可以发现无论是在数据库中还是在现实生活中，类别之间都是有关系的。相似的类别之间分享相同的视觉信息特征和运动轨迹特征。所以独立为每个类别训练分类器会浪费掉它们之间共享的信息。另一方面，仅有的几种多任务行为识别方法认为所有行为之间都共享同一个特征空间，这个假设对绝大部分情况都不太现实，可能会导致“负迁移”损害模型性能。通过观察可以发现，数据库中的行为类别的关系有亲疏之分，相似程度高的行为可以分为一组，共享相同的特征空间。本章的研究的重点在于挖掘数据库中行为类别之间潜在的结构关系，提出了基于SVM间隔的多任务行为识别。具体来说，首先对于数据库中的每对行为训练一个$1-vs-1$的SVM，通过利用得到的支持向量等结果计算两个行为之间的SVM间隔，并用其来衡量行为之间的相似度。SVM间隔越小，说明行为之间相似度越高，分享的共同信息越多，反之亦然。随后将所有类别对应的相似度矩阵利用反向传播算子聚类法可以得到类别中潜在的结构关系。通过将此结构关系作为先验正则项与多任务学习框架结合，可以鼓励存在于一组内的行为类别共享特征，而不在一组内的行为类别之间进行特征竞争。通过在HMDB51数据库上进行实验，可以表明基于SVM间隔的多任务行为识别较之其他的多任务行为识别方法的优越性，不仅仅如此，本章的方法与行为识别领域一些先进的方法比较也有了性能上的提高。这都表明了，利用行为类别之间的信息，寻找行为类别之间潜在的行为结构是必要的。

第四章　基于互信息的多任务行为识别

4.1　引言

如本文在第三章所述，目前现存的行为识别方法的一般流程都是提取视频特征再利用特征训练分类器。它们之中的绝大多数将每个行为类别视为一个任务，然后独立的对每个任务学习其对应的分类模型，从而浪费和忽视了类别之间共享的信息。基于此，第三章中提出了基于SVM间隔的多任务行为识别，首先对数据库中的所有任务两两训练1对1的SVM，然后用SVM的间隔作为对应的两个类别之间的相似性度量。通过反向传播算子法对相似度矩阵进行聚类，可以得到数据库的潜在的类别结构，最后将此结构作为先验约束与多任务学习框架结合，从而利用任务之间的相关信息提高分类模型的性能和泛化能力。

本章内容在第三章的基础上，对于任务之间的相似度度量做出了更深层次的探索。既然多任务学习的目的是令相似的任务之间分享特征，那么如果从特征的角度来衡量任务之间的相似度将更加的直观和合理。本文使用的特征为Fisher Vector[22][23][24]，它近年来被行为识别方法广泛利用并取得了很好的效果。Fisher Vector由高斯混合模型的最大似然函数分别对于均值和协方差的导数串联而成，也就是说Fisher Vector中的连续维度对应着一个特定的高斯函数。此外，我们观察到高斯混合模型中每个高斯函数对于不同行为类别的贡献不同，不同的高斯混合模型都捕捉和表示了不同的视觉特征和运动特征。在这些观察的基础上，我们提出了一个假设：相似的行为类别分享较多的高斯函数，也就是分享对应的Fisher Vector中的对应维度的特征。基于此，我们提出了基于互信息的多任务行为识别。首先我们计算出每个高斯函数对于每一个行为类别的互信息来衡量其对这个类别的重要性权重，如果这个权重过低，那么这个高斯函数对应的特征也被视为无用的。两个类别同时拥有的高斯函数的个数用于衡量类别之间的相似度。通过计算得到所有类别的相似度矩阵之后，再如第三章所说的，通过聚类得到类别之间隐含的特征结构，最后利用基于组结构约束的多任务学习为每个组学习到一个共享的特征空间，具体的流程框架图如图4.1所示。

综上所述，本章提出了一种互信息的多任务行为识别方法。本章的主要内容安排如下：在4.2节中介绍Fisher Vector；在4.3节中介绍基于互信息的多任务行为识别，首先会介绍基于互信息的相似性度量，然后简单介绍目标函数的优化；在4.4节中进行实验仿真并对实验结果进行分析；在4.5节中对本章内容进行总结。

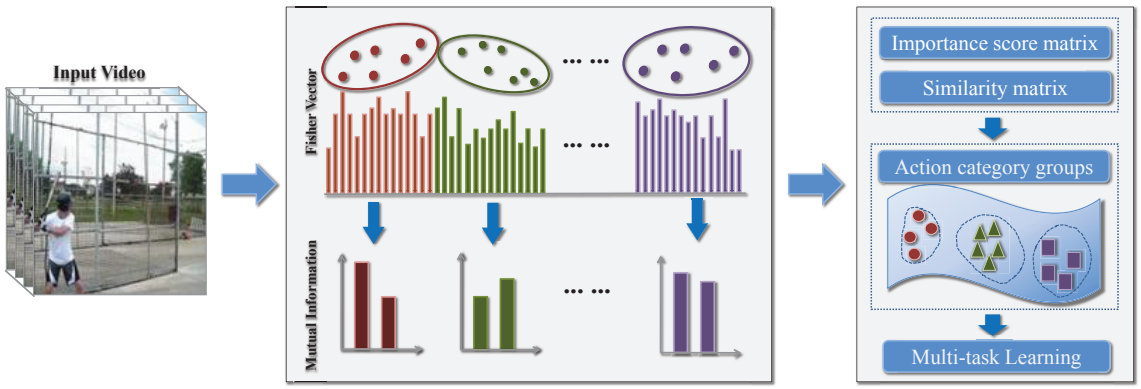


图 4.1 基于互信息的多任务行为识别的具体流程。

4.2 Fisher Vector

词袋模型是计算机视觉领域常用的特征处理方法，它可以把从视频和图像中提取的局部特征转化为全局特征。词袋模型的具体步骤如图1.1所示，分别为（1）特征提取，（2）特征预处理，（3）生成码本，（4）特征编码，（5）特征的池化和正规化。这五个步骤中的每一个都有研究者们提出了许多有效的方法，而其中最重要的当数码本的生成和特征的编码。如果使用高斯混合模型（Gaussin Mixture Model）来生成码本和基于Fisher Vector 的方法对特征进行编码，那么最终得到的特征就称为Fisher Vector。Fisher Vector[47] 起源于Fisher Kernel。通过将原始特征通过Fisher Kernel映射到高维的特征空间之后得到的向量即为Fisher Vector。下面将分别对Fisher Kernel 和Fisher Vector 做出简单的介绍。

Fisher Kernel[?][48]是分类问题中常用的模型，它结合了生产模型和判别模型的优点。假定给定了样本X，它的生成过程可以被建模为一个包含参数λ 的概率密度函数p。在我们的应用中，X为一个视频样本。则X可以被如下的梯度向量所描述：

$$G_{\lambda}^{X} = \nabla_{\lambda} \log p(X \mid \lambda), \tag{4-1}$$

上式中的似然函数的梯度表示了参数λ对于X生成过程的贡献。这个梯度向量的维度仅仅和λ向量的维度相关。基于上述梯度向量，定义核函数

$$K(X,Y) = G_{\lambda}^{X'} F_{\lambda}^{-1} G_{\lambda}^{Y}, \tag{4-2}$$

其中F_{λ}为p的Fihser信息矩阵：

$$F_{\lambda} = E_{x \sim p}\left\langle \nabla_{\lambda} \log p(x \mid \lambda) \nabla_{\lambda} \log p(x \mid \lambda)' \right\rangle, \tag{4-3}$$

其中F_λ是对称且正定的矩阵，它的柯勒斯基（Cholesy)分解为$F_\lambda = F_\lambda - F_\lambda$，则可得正规化向量：

$$g_\lambda^X = L_\lambda G_\lambda^X, \tag{4-4}$$

公式(4-2)可被重写为g_λ^X的点乘形式，其中g_λ^X即为样本X的Fisher Vector。从上面的分析可以看出将原始样本映射到特征空间上得到的特征向量，就是最终的Fisher Vector。

假设样本$X = \{x_t, t = 1, 2, ..., T\}$表示从视频中提取的描述子，其中$T$为描述子的个数，而$X$的生成过程可由$p$所独立描述。$p$为概率密度函数，它可以有很多的形式，通常将它选择为高斯混合模型:

$$p(x) = \sum_{i=1}^{N} w_i p_i(x), \tag{4-5}$$

高斯混合模型中的每一个元素p_i都是独立的高斯函数，它可以视为词袋模型码本中的一个单词（word），每个p_i都表示了不同的视觉特征和运动特征，N表示了码本的大小。在下面的文章中，我们统一将码本中的词称为高斯函数。令参数$\lambda = \{w_i, \mu_i, \sum_i, i = 1, 2, ..., N\}$，其中$w_i, \mu_i$ 和$\sum_i$分别表示第i 个高斯函数的权重，均值和协方差均值。假设协方差矩阵$\sum_i$为对角矩阵，则可以方差向量的平方σ_i^2表示协方差矩阵。则可以用最大似然函数估计（Maximum Likelihood Estimation）在一系列样本的基础上训练得到高斯混合模型p，通常使用EM 算法迭代解决这个问题。

高斯混合模型和K-means算法一样，是词袋模型中生成码本的一种方法。高斯混合模型与K-means方法的不同之处在于，后者是一种硬聚类方法，即它必须将特征描述子划分给码本中确切的一个词，但是前者可以通过EM算法给出描述子属于码本中每个词的可能性，所以高斯混合模型是一种软聚类方法，它比K-means算法更加的灵活，它不仅仅描述了码本中词的信息，而且还描述了码本分布情况。

已知Fisher Vector由公式(**??**)的似然函数的梯度向量对于参数λ的导数串联而成。将高斯混合模型带入公式(**??**)可得

$$\begin{aligned} \boldsymbol{\rho}_i &= \frac{1}{\sqrt{\pi_i}} \gamma_i \left(\frac{\mathbf{x} - \boldsymbol{\mu}_\mathbf{i}}{\boldsymbol{\sigma}_i} \right), \\ \boldsymbol{\tau}_i &= \frac{1}{\sqrt{2\pi_i}} \gamma_i \left[\frac{(\mathbf{x} - \boldsymbol{\mu}_\mathbf{i})^\mathbf{2}}{\boldsymbol{\sigma}_\boldsymbol{i}^\mathbf{2}} - 1 \right], \end{aligned} \tag{4-6}$$

其中$\gamma_k = p(i \mid x_t)$表示局部描述子属于第i个高斯函数的权重：

$$\gamma_k = \frac{\pi_k \mathcal{N}(x; \mu_k, \sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x; \mu_i, \sigma_i)}, \tag{4-7}$$

$\boldsymbol{\rho}_i$和$\boldsymbol{\tau}_i$表示第i个高斯函数分别对于均值μ_i和协方差矩阵σ_i^2 的导数。这两个向量的维数和描述子的维数相同。将它们串联起来可得

$$g_\lambda^X = [\rho_1, \tau_1, ..., \rho_N, \tau_N]. \tag{4-8}$$

g_λ^X这个$2ND$维度的向量即为描述X的Fisher Vector。

Fisher Vector相对于传统的词袋方法的优势在于，后者得到的是一个极其稀疏的向量，它只关注了关键词的数量信息，所以是一个0阶的统计信息；Fisher Vector并不稀疏，同时，除了描述子的0阶信息，Fisher Vector还包含了1阶(期望)信息、2阶(方差信息)，因此Fisher Vector是一个高维的向量特征。由于Fisher Vector保留了码本中每个词更丰富的信息，所以码本的大小也比传统的词袋方法要小。终上所述，Fisher Vector是一个高维且包含了更多信息的特征。

4.3 基于互信息的多任务行为识别

在上一小节中，我们介绍了Fisher Vector，知道了它事实上由高斯混合模型的最大似然函数分别对于均值和协方差的导数串联而成，如图4.1所示Fisher Vector中的连续维度特征对应着一个特定的高斯函数。基于每个高斯函数都捕捉了视频的不同特征和相似的视频类别之间又共享相同的特征这两个假设，我们提出了基于互信息的多任务行为识别，首先利用互信息来衡量行为类别之间的相似度并探索了数据库中潜藏的类别机构信息，最后将此结构信息作为多任务学习的先验正则项来识别行为。

4.3.1 基于互信息的相似性度量

本小节主要描述了一种基于互信息的衡量两个行为类别之间相似度的方法。首先视频都先被表示为Fisher Vector，令每个高斯函数对于每个行为类别的互信息作为该高斯对于这个类别的权重，权重越大说明该高斯函数对应的Fisher Vector对于该行为类别就越重要，而权重小的高斯函数对应的特征应该在多任务学习之后被抛弃。两个类别共同持有的高斯函数的个数用于衡量两个类别之间的相似度。

假设给定数据库中共有C个行为类别，对于第c个类，对应的训练数据为$\{\mathbf{x}_{c,i}, y_{c,i}\}_{i=1}^{N_c} \subset \mathrm{R}^D(c = 1, .., C)$, i和N_c 分别表示第c个行为类别中训练样本的序号和个数。通过将K个高斯函数对应的梯度向量串联可将每个是样本表示为$\{\mathbf{x}_{c,i}^k\}_{k=1}^K$，其中$k$ 表示高斯混合模型中高斯函数的序号，$\mathbf{x}_{c,i}^k$为第k个高斯函数对应的特征向量。

互信息（Mutual Information，MI）是概率论和信息论中常用的一个量度，它一般用来衡量变量间相互依赖性的量度。在特征选择应用中，互信息被用来衡量某个特征和特定类别的相关性，如果信息量越大，那么特征和这个类别的相关性

越大。反之也是成立的。基于此理论基础，为了探求高斯函数与行为类别之间的关系，我们提出了利用互信息的来衡量一个高斯函数对于一个行为类别的重要性。具体来说，令$I_c(\mathbf{x}_{:}^{k}, \mathbf{y})$表示第$c$个行为类别和第$k$个高斯函数之间的互信息，其中$\mathbf{x}_{:}^{k} = \{\{\mathbf{x}_{c,i}^{k}\}_{i=1}^{N_c}\}_{c=1}^{C}$为将数据库的所有视频表示为Fisher Vector之后，第k个高斯函数对应的特征维度组成的矩阵，其中下标:表示了它包含来自所有样本的特征。$\mathbf{y}$ 表示类标空间，互信息的值可由如下公式计算：

$$I_c(\mathbf{x}_{:}^{k}, \mathbf{y}) = H(\mathbf{y}) + H(\mathbf{x}_{:}^{k}) - H(\mathbf{x}_{:}^{k}, \mathbf{y}), \tag{4-9}$$

上式中H表示一个随机变量的熵，是随机变量不确定性的度量。y的值随类别标签c的值变化而变化。当计算I_c时，所有属于第c个类别的视频样本都是视为正样本，则对应的y 值为1，而其他不属于第c个类别的所有样本都视为负样本，y的值为-1。这个步骤的处理不仅仅方便了互信息的计算，而且和最终对每个类别训练二元线性分类器的时候是一致的。在最终的多任务学习模型中，在对每个类别训练分类器的时候对类标的处理也是相同的方式。

如果是设X是一个取有限个值的离散随机变量，其概率分布为

$$P(X = x_i) = p_i, i = 1, 2, .., n \tag{4-10}$$

则随机变量X的熵H通常被如下的公式所计算：

$$H(X) = -\sum_{i} P(x_i) \log_b P(x_i), \tag{4-11}$$

上式中b表示对数的基数。H只依赖于X的分布，而与X的取值无关。在我们的工作中X表示Fisher Vector。如果要通过计算概率分布函数来求解$H(\mathbf{x}_{:}^{k})$将会耗费大量的时间和精力，所以我们采取了[49]的方法，利用了频率来代替概率函数。具体的，我们如公式 (4-13)所示将Fisher Vector的中的值量化为若干个组，为了计算的简单，在我们的实验中我们将组的个数设置为了2。其他的量化方法也是可行的，通过实验我们发现组的个数对实验结果并没有太大的影响。在将视频表示为Fisher Vector之后，通过统计每个高斯函数对应的Fisher Vector特征值的分布规律，可知0必须被设为划分组的阈值，具体分布如图4.2所示。

$$x = \begin{cases} 0 & x < 0 \\ 1 & otherwise, \end{cases} \tag{4-12}$$

通过对Fisher Vector做了量化处理之后，则公式 (4-9)中的离散熵可以被如下公式

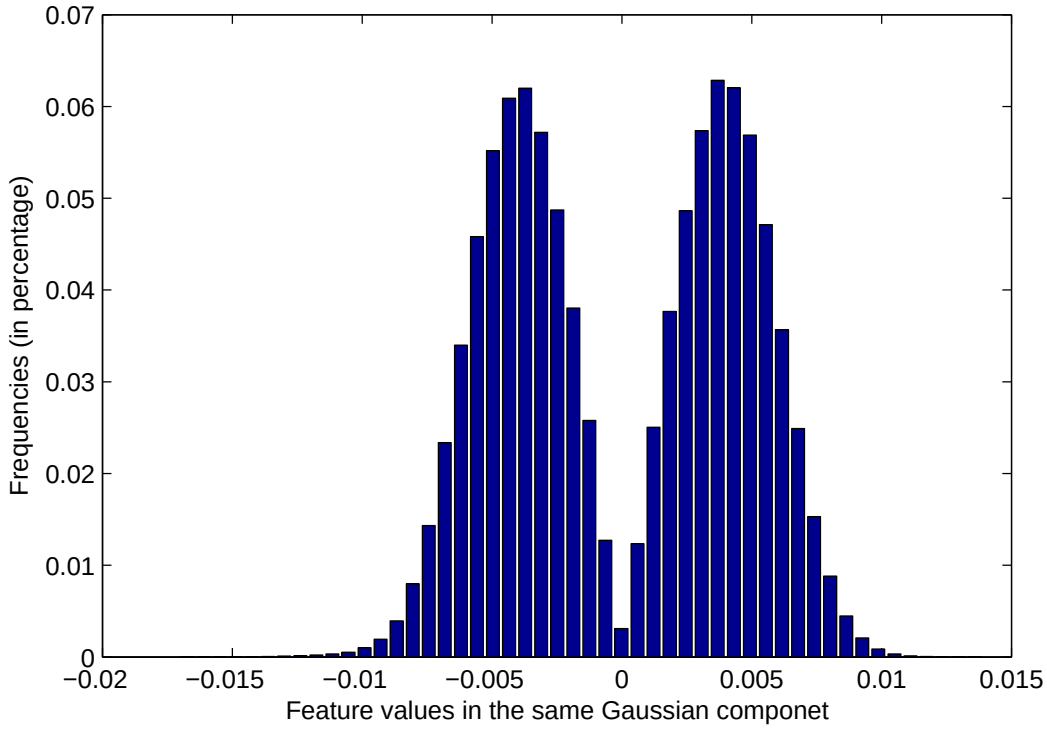


图 4.2 Fisher Vector中同属于一个高斯函数的特征值直方图

所计算：

$$H(\mathbf{x}_{:}^{k}) = -\sum_{j=1}^{2} p_j \log_2(p_j), \tag{4-13}$$

上式中p_j表示的是矩阵$x_{:}^{k}$中的每一项落入第j个组的频率。所有的p_j都是可以通过得到数据库的Fisher Vector之后，通过统计其特征值分布统计计算得到的。

在[49]的工作中，Yu Zhang等人提出了基于互信息的排序方法，并应用于特征的选择和对特征降维。我们的工作和他们尽管都用到了互信息，但是在细节和规划上却有很大的不同。前者仅仅计算了每一单独维度的特征和图像类别之间的互信息，而忽略了每个维度的特征对于不同类别的重要程度不同的事实。我们计算的是一个高斯函数对应的连续维度的特征对于行为类别的互信息，从而衡量此高斯函数对于类别的重要性。我们的目的是不仅仅可以为每一个行为类别挑选出最具有判别性的特征从而训练出可信赖的二元线性分类器，而且可以通过量化类别持有的高斯函数的个数来衡量类别之间的相似性。从而探索数据库潜藏的类别结构信息。此外，Yu Zhang将互信息低的特征维度直接抛弃达到特征的选择和降维，而我们利用了多任务学习方法和组结构约束来实现特征选择。

当我们为所有的高斯函数分别计算出它们对于每个行为类别的互信息I_c之后，

我们用此来衡量高斯函数对于行为类别的重要程度，I_c越大，说明该高斯函数对应的Fisher Vector特征维度对于该行为类别越重要。为了衡量行为类别之间的相似性，我们利用阈值λ将正规化之后的互信息矩阵I转化为一个值为0或者1的二元矩阵，具体公式如下所示：

$$I_{i,j} = \begin{cases} 0 & I_{i,j} < \lambda \\ 1 & otherwise, \end{cases} \tag{4-14}$$

$I \subset \mathrm{R}^{C \times K}$表示了对于所有行为类别而言高斯函数的重要性权重。上式说明了对于行为类别不重要的特征应该抛弃，重要的应该保留。所以利用阈值$lambda$ 来量化互信息矩阵I。$I_{i,j}$如果大于λ则令其为1，说明对于第i个行为类别，第j个高斯函数对应的Fisher Vector很重要，故予以保留，反之抛弃。如果行为类别同时保留的高斯函数的个数越多，则说明它们之间的相似度越高。

将互信息矩阵I转化为二元矩阵之后，就可以基于它来计算相似度矩阵$S \subset \mathrm{R}^{C \times C}$，我们用$s_{i,j}$来表示第$i$个类别和第$j$个类别之间的相似度，它可由如下的公式计算：

$$s_{i,j} = \sum(\mathbf{I}_i \odot \mathbf{I}_j), \tag{4-15}$$

上式中$\odot$表示按位与，即将第i个类别与第j个类别的高斯权重行向量按位与之后再求和，所得的值正是两个类别同时选择的类别的个数，也就是它们之间的相似度$s_{i,j}$。通过计算互信息矩阵I中的每两行之间按位与之和后，可以得到一个对称的相似性矩阵S。

可以发现此相似性度量方法是直接从特征的角度出发，所以它得到的类别之间的相似性关系也是基于特征的。具体而言，相似度矩阵S和互信息I都是基于通过估计特征$\mathbf{x}$的概率分布函数计算得到的。所以如果使用不同的特征描述子，由于它们表示了不同的特征空间和分布，从而可能使得类别之间的相似结构也是不同的。这也是基于互信息相似性度量方法的灵活性和优越性。

4.3.2　基于互信息的多任务行为识别

基于互信息的多任务行为识别方法的具体流程如图4.1所示。通过对Fisher Vector向量和数据库中行为类别的观察。我们得到了两个重要假设：（1）Fisher Vector中的连续维度的特征对应着一个高斯函数，不同的高斯函数捕捉了视频特定的视觉和运动特征，所以高斯函数对于行为类别的重要程度是不同的。（2）数据库中的行为可以分为若干个组，每个组内的任务都是相似的且共享相同的特征空间，即会共享更多的来自于同一个高斯函数的特征维度。基于这两个重要假设，我们首先将视频表示为Fisher Vector，然后利用4.3.1 中的方法通过计算高斯函数对于行为类别的互信息得到类别之间的相似度矩阵S。然后和第三章的方法类似，通过反向传播

算子聚类法得到基于训练特征的最优的任务结构之后，利用基于组结构约束的多任务学习框架得到分类模型。

具体的目标函数如 公式(**??**)所示，区别在于在本方法中先验结构约束的计算方法是不同的。基于SVM间隔的行为识别利用行为之间的SVM间隔来衡量行为之间的相似度，而本方法通过计算高斯函数对于行为类别的互信息得到类别之间的相似度矩阵S。

优化求解公式(**??**)之后，对于新的测试视频，将它表示为Fisher Vector$x_n \in \mathbf{R}^D$之后，它对应的标签y_n可由公式 (**??**)求得。

4.4 实验结果与分析

本小节先介绍了实验的具体细节，包括数据库、特征、基准方法和参数。然后展示了基于互信息的多任务行为识别和对比方法分别在行为识别数据库HMDB51和UCF50上的结果，并给出了结果的分析。

4.4.1 实验设计

本方法的实验步骤与基于SVM间隔的行为识别方法的实验步骤非常相似，所以在此不再对特征、基准方法和参数选择上再做赘述，具体可以参照3.4.1小节，下面只简单的对数据库做介绍。

本实验中使用的数据库之一UCF50[50]总共高寒6617个视频，它们分别啦子50 个行为类别。如图4.3所示所有的视频都来自于Youtube。我们的对于训练数据、验证数据还有测试数据的划分都依照了[45][35]中的步骤。此外，我们还依照了[35]的方法，分别用25%, 50%, 75% and 100%的训练数据进行了实验。每次设置下都将实验重复了十次。UCF50的部分视频样本示例如图4.3所示。

HMDB51数据库在第三种中已做了介绍，除了基本的实验之外，为了展示多任务学习的优点，我们同样随机选择了10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%的训练数据进行了实验，如上所述，每次设置下都将实验重复了十次。

我们选择MAP(mean average precision)作为最终的评估参数。

4.4.2 实验结果与分析

首先如表4.1所示，我们比较了基于互信息的多任务行为识别与三个基准方法的识别精度。从表中的结果可以看出，基于互信息的多任务行为识别不管是利用哪个类型的描述子，与基准方法相比识别精度都是最高的。我们的多任务学习框架利用了组结构约束，它综合了ℓ_1 范数和ℓ_{21}范数的优点。当所有的行为类别都属于同一个组的时候，也就是所有的类别都相似的情况下，公式(**??**) 中的结构正则项退化为ℓ_{21}

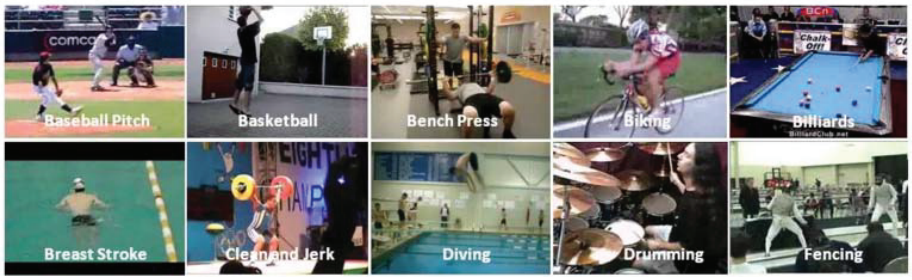


图 4.3　UCF50中的样本示例。

特征/方法	多任务行为识别	基于ℓ_1范数的多任务行为识别	基于ℓ_{21}范数的多任务行为识别	基于互信息的多任务行为识别
HOG	40.22	44.05	43.84	**45.92**
HOF	49.00	50.04	50.09	**50.98**
MBHx	40.24	42.81	42.61	**44.03**
MBHy	47.13	48.60	48.62	**49.57**
combined	60.15	60.38	60.31	**60.84**

表 4.1　基于互信息的多任务行为识别与基准方法在HMDB51上的实验结果对比。

范数。当所有的行为类别都自成一组时，也就是所有的类别之间都不共享特征的情况下，结构正则项退化为ℓ_1范数。

基于互信息的多任务行为识别和单任务学习在训练样本变化时的实验结果如表4.2所示。很明显，当训练数据较少的时候多任务学习与单任务学习相比得到了很大的提升。这个提升是由于多任务学习的特征共享机制导致了。这个结果也说明了我们关于数据库中的行为可以分为若干个组，每个组内的任务都是相似的且共享相同的特征空间这个假设的正确性。

最后我们把基于互信息的多任务行为识别和行为识别领域一些先进的方法在HMDB51数据库上做了对比试验。如表3.2所示，基于SVM间隔的多任务行为识别甚至比基于深度学习的方法[44]效果还要好。

然后我们在UCF50上进行了实验对比。如图4.5所示，我们展示了一部分

Percentage	10%	20%	30%	40%	50%
Single Task	36.17 ± 1.79	45.45 ± 1.30	50.39 ± 1.15	53.17 ± 0.88	54.72 ± 0.71
Proposed Work	40.03 ± 2.11	48.10 ± 1.75	52.39 ± 1.43	55.03 ± 1.12	56.53 ± 0.94
Percentage	60%	70%	80%	90%	100%
Single Task	56.45 ± 0.73	57.68 ± 0.53	58.61 ± 0.65	59.54 ± 0.40	60.22
Proposed Work	57.20 ± 0.88	59.15 ± 0.81	59.48± 0.73	60.14 ± 0.52	60.84

表 4.2　基于互信息的多任务行为识别和单任务学习在训练样本变化时在HMDB51数据库的实验结果对比，该实验中用的特征为四种描述子的串联向量。

方法	精度
Sadanand and J.Corso[45]	26.9%
Yang et al.[46]	53.9%
Hou et al.[40]	57.88%
K. Simonyan and A. Zisserman[44]	59.5%
基于互信息的多任务行为识别	**60.84**%

表 4.3 基于互信息的多任务行为识别方法与行为识别领域一些先进方法在HMDB51上的实验结果对比。

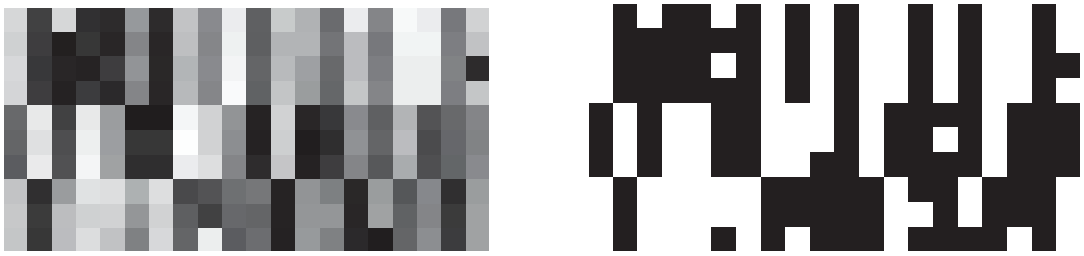


图 4.4 **Left:** 在UCF50数据库上，基于HOG特征训练计算得到的互信息矩阵I的一部分。行和列分别表示了行为类别和高斯函数。矩阵I中颜色的深浅代表了对应高斯函数对于行为类别的重要程度。**Right:** 通过I计算得到的相似度矩阵S。白色表示0。

在UCF50数据库上，基于HOG特征训练计算得到的互信息矩阵I和它对应的相似度矩阵S。在相似性矩阵I中，行代表了行为类别，列代表了高斯函数。$I_{i,j}$的深浅表示了对于第i个行为类别而言，第j个高斯函数的重要程度。通过对I进行公式(4-15)的处理得到矩阵S，$s_{i,j}$是白色表示对于第i 个类而言，第j个高斯函数对应的特征维度不太重要，所以抛弃它。从图4.5中很容易可以观察到，同属于一组的行为类别更容易同时和一个高斯函数具有更密切的关系。当然S 和I也存在一些异常任务，这说明了组结构约束只是一个软约束，它非常的灵活。

如表4.4所示，我们在数据库UCF50上比较了基于互信息的多任务行为识别与三个基准方法的识别精度。从表中的结果可以看出，基于互信息的多任务行为识别不管是利用哪个类型的描述子，与基准方法相比识别精度都是最高的。这个结果和我

特征/方法	多任务行为识别	基于ℓ_1范数的多任务行为识别	基于ℓ_{21}范数的多任务行为识别	基于互信息的多任务行为识别
HOG	82.37	83.4	83.42	**84.04**
HOF	85.69	87.69	86.89	**87.98**
MBHx	83.18	83.94	84.03	**84.54**
MBHy	85.97	87.08	86.82	**88.02**
combined	88.13	88.69	88.54	**89.17**

表 4.4 基于互信息的多任务行为识别与基准方法在UCF50上的实验结果对比。

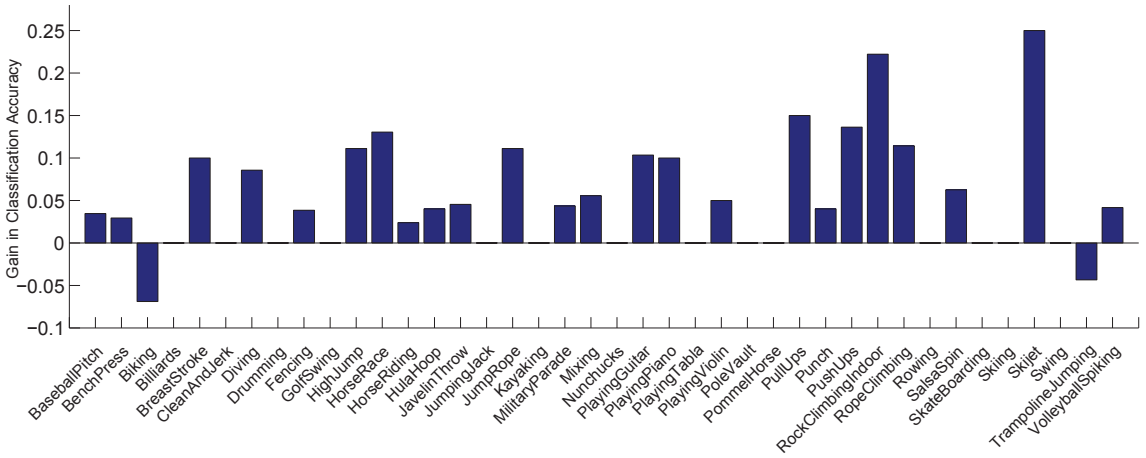


图 4.5　当仅使用25 %的训练数据时，基于互信息的多任务行为识别方法在UCF50行为数据库中的绝大多数类别上的识别精度的提升。

Methods	Accuracy
Sadanand and J.Corso[45]	57.9 %
Kishore K. Reddy and Mubarak Shah[50]	76.9%
Zhou et al.[35]	78.3%
Wang et al.[16]	85.9%
Proposed Method	**89.2**%

表 4.5　基于互信息的多任务行为识别与基准方法在UCF50上的实验结果对比。

们在HMDB51上的结果一致，原因也是如上文中分析的：我们的多任务学习框架利用了组结构约束，它综合了ℓ_1 范数和ℓ_{21} 范数的优点。图4.4展示了当仅使用25 %的训练数据时，基于互信息的多任务行为识别方法在UCF50行为数据库中的绝大多数类别上的识别精度的提升。值得注意的是，同属于一组的行为类别普遍获得了显著的识别精度上的提升，，这都归功于多任务学习框架利用了原本被忽略了类别之间的信息。

在表4.5中，我们把基于互信息的多任务行为识别和行为识别领域一些先进的方法在UCF50数据库上做了对比试验。在这些方法中，Qiang Zhou[35]等人的工作和我们提出的方法略为相似。他们的工作本文在2.3.2小节中已做了简单的介绍。Qiang Zhou提出了一种基于多任务学习框架的行为识别方法，他们认为任务之间分享的基任务可以视为不同的行为类别之间共享的运动元素。比如跑步、走路、挥手等运动其实都是有身体部位的基本移动构成的。最终每个类别的分类器参数由基本运动元素的线性组合而成。和他们的工作相比，我们的方法首先认为行为类别之间共享的是确切对应着同一个高斯函数的的连续维度的特征，其次我们通过互信息来寻找潜藏在数据库中的行为结构，最后在Qiang Zhou的工作中共有3个参数需要调节，而且基任务的个数需要事先确定，而基于互信息的多任务行为识别的方法参数的个数比

较少，而且由于反向传播算子聚类法的使用，行为可分为的组的个数也不用事先确定。

4.5 本章小结

在第三章中提出的基于SVM间隔的多任务学习方法通过利用SVM间隔计算行为类别的相似性矩阵，然后聚类得到数据库的潜在的类别结构，最后将此结构作为先验约束与多任务学习框架结合，通过利用任务之间的相关信息提高分类模型的性能和泛化能力。本文通过观察常用特征Fisher Vector，得到了以下假设：（1）Fisher Vector中的连续维度的特征对应着一个高斯函数，不同的高斯函数捕捉了视频特定的视觉和运动特征，所以高斯函数对于行为类别的重要程度是不同的。（2）数据库中的行为可以分为若干个组，每个组内的任务都是相似的且共享相同的特征空间，即会共享更多的来自于同一个高斯函数的特征维度。基于这两个假设，本文基于SVM间隔的多任务行为识别方法提出了一种新的用来衡量行为类别之间相似度的方法，即基于互信息的相似性度量。通过计算出每个高斯函数对于每一个行为类别的互信息来衡量其对这个类别的重要性。得到数据库的相似性矩阵之后，利用反向传播算子聚类法得到类别中潜在的结构关系。基于互信息的多任务行为识别方法可以鼓励同属于一组的任务共享了来自于一个高斯函数对应的特征，它直接从特征的角度来计算行为类比的相似性，与基于SVM间隔的多任务行为识别相比更加的直观，计算也更加的简单。在数据库HMDB51和数据库UCF50上的实验实验结果都表明基于互信息的多任务行为识别较之其他的多任务行为识别方法的优越性。

第五章　总结与展望

5.1　总结

近年来，随着互联网的不断普及，Youtube、优酷土豆等视频社交网站的出现，拍摄视频成本的持续降低，使得互联网上的视频数据的数量变的越来越庞大。视频作为一种信息载体，与传统图片、声音媒体介质相比信息传播的效率更高，内容也更加的丰富。基于视频的人体行为识别这一需求也应运而生。识别视频中的人体做的行为具有广阔的应用市场，例如智能人机交互、视频监控、基于内容的视频检测等。随着视频数量的指数级增长，一个精确又强健的人体行为识别方法变得至关重要。然而，不幸的是，由于视频本身包含的场景和环境信息十分复杂，且训练样本过少等原因，目前行为识别方法的应用仅仅停留在实验室中，无法满足工业界的要求。另一方面，现存的行为识别方法绝大多数都单独的为每一个行为类别训练其对应的分类模型，从而忽略了类别之间的信息。事实上，行为类别不是独立无关联的，它们之间分享了很多的信息，通过利用这些信息可以有效的减少对训练样本的依赖和解决视频中由于遮挡产生的问题。基于此，本文从多任务学习的角度来研究了行为识别，我们通过观察发现数据库中的行为并不是独立的，而是存在潜在的结构信息。数据库中的行为可以被分为若干个组，位于同一个组内的行为是相似的，也共享一个特征空间。如何得到这个潜在的结构，正是本文研究的重点。本文提出了两种衡量行为类别之间相似性的方法，并与多任务学习相结合，主要的研究成果如下：

（1）提出了一种基于SVM间隔的多任务行为识别方法。具体来说，首先对于数据库中的每对行为训练一个$1-vs-1$的SVM，通过利用得到的支持向量等结果计算两个行为之间的SVM 间隔，并用其来衡量行为之间的相似度。SVM间隔越小，说明行为之间相似度越高，分享的共同信息越多，反之亦然。随后将所有类别对应的相似度矩阵利用反向传播算子聚类法可以得到类别中潜在的结构关系。通过将此结构关系作为先验正则项与多任务学习框架结合，可以鼓励存在于一组内的行为类别共享特征，而不在一组内的行为类别之间进行特征竞争。通过在HMDB51数据库上进行实验，可以表明基于SVM间隔的多任务行为识别较之其他的多任务行为识别方法的优越性，不仅仅如此，该方法与行为识别领域一些先进的方法比较也有了性能上的提高。这都表明了，利用行为类别之间的信息，寻找行为类别之间潜在的行为结构是必要的。

（2）提出了一种基于互信息的多任务行为识别方法。该方法在基于SVM间隔

的多任务行为识别方法的基础上，提出了一种新的用来衡量行为类别之间相似度的方法，即基于互信息的相似性度量。通过计算出每个高斯函数对于每一个行为类别的互信息来衡量其对这个类别的重要性。得到数据库的相似性矩阵之后，利用反向传播算子聚类法得到类别中潜在的结构关系。基于互信息的多任务行为识别方法可以鼓励同属于一组的任务共享了来自于一个高斯函数对应的特征，它直接从特征的角度来计算行为类比的相似性，与基于SVM间隔的多任务行为识别相比更加的直观，计算也更加的简单。在数据库HMDB51和数据库UCF50上的实验实验结果都表明基于互信息的多任务行为识别较之其他的多任务行为识别方法的优越性。

5.2 展望

尽管海内外有很多学者针对人体行为识别做出了很多研究，然而该领域仍有很多技术难题没有得到解决。本论文通过观察，针对如何寻找和利用行为数据库中潜在的行为结构做出了研究。提出了两种衡量类别之间相似性从而将类别分组的方法，与基于组结构正则项的多任务学习结合完成行为的识别。数据库上的实验结果验证了我们这两种方法的有效性。尽管本文提出的方法在一定的程度上利用到以往被忽视的结构信息，改善了人体行为识别的性能。但是还是有许多实际的问题需进一步的深入和完善。结合本文的技术基础，我们列举了下一步的研究方向：

（1）寻找更好的视频的特征表示。在第一章中提到，目前描述视频最好的特征为密集轨迹，然而提取时间过长，计算量太大等缺点。基于密集轨迹的行为识别的实验结果说明了，特征是一个行为识别方法是否可以推广的一个决定性因素。如果一个特征可以很好的刻画视频中行为的视觉和运动信息，那么它就可以推广到大部分的应用中。

（2）寻找更好的衡量行为相似度的测量。如果想利用行为之间的信息，不可避免的要衡量行为之间的关系。然而目前在这方面的研究和方法非常少。尽管本文提出了两种计算行为之间相似度的方法，但是仍然比较初级且有局限性。在今后的研究中，如何寻求更好的行为之间相似性距离测度，将会是一个很有价值的研究方向。

（3）寻找更好的多任务学习正则项。多任务学习的正则项决定了多任务学习将如何利用行为之间共享的信息。正确的对行为特征空间的假设和建立会更好的利用特征之间的信息。目前已存在很多多任务学习方法，但是针对人体行为的还较少。在今后的研究中，研究行为之间的关系，设计符合行为结构的正则项，将会有效的帮助行为的有效识别。

参考文献

[1] AGGARWAL J K, RYOO M S. Human activity analysis: A review.[J]. ACM Comput. Surv., 2011, 43(3) : 16.

[2] KE S-R, HOANG L U T, LEE Y-J, et al. A Review on Video-Based Human Activity Recognition.[J]. Computers, 2013, 2(2) : 88 – 131.

[3] COLLINS R T, LIPTON A J, KANADE T, et al. A System for Video Surveillance and Monitoring[J], 2000.

[4] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as Space-Time Shapes[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2007, 29(12) : 2247 – 2253.

[5] DARRELL T, PENTLAND A. Space-time gestures[C] // CVPR. 1993 : 335 – 340.

[6] BERNDT D J, CLIFFORD J. Using Dynamic Time Warping to Find Patterns in Time Series[C] // Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop, Seattle, Washington, July 1994. Technical Report WS-94-03. 1994 : 359 – 370.

[7] GREEN R D, GUAN L. Quantifying and recognizing human movement patterns from monocular video Images-part I: a new framework for modeling human motion[J]. IEEE Trans. Circuits Syst. Video Techn., 2004, 14(2) : 179 – 190.

[8] BOBICK A F, DAVIS J W. The Recognition of Human Movement Using Temporal Templates[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2001, 23(3) : 257 – 267.

[9] KE Y, SUKTHANKAR R, HEBERT M. Spatio-temporal Shape and Flow Correlation for Action Recognition[C] // CVPR. 2007.

[10] RODRIGUEZ M D, AHMED J, SHAH M. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition[C] // CVPR. 2008.

[11] BLANK M, GORELICK L, SHECHTMAN E, et al. Actions as Space-Time Shapes[C] // ICCV. 2005.

[12] YILMAZ A, SHAH M. Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras[C] // ICCV. 2005 : 150 – 157.

[13] SHEIKH Y, SHEIKH M, SHAH M. Exploring the Space of a Human Action[C] // ICCV. 2005 : 144 – 149.

[14] RAO C, SHAH M. View-Invariance in Action Recognition[C] // CVPR. 2001 : 316 – 322.

[15] WANG H, KLÄSER A, SCHMID C, et al. Action Recognition by Dense Trajectories[C] // CVPR. 2011.

[16] WANG H, SCHMID C. Action Recognition with Improved Trajectories[C] // ICCV. 2013.

[17] CHOMAT O, CROWLEY J L. Probabilistic Recognition of Activity using Local Appearance[C] // CVPR. 1999 : 2104 – 2109.

[18] ZELNIK-MANOR L, IRANI M. Event-Based Analysis of Video[C] // CVPR. 2001 : 123 – 130.

[19] LAPTEV I, LINDEBERG T. Space-time Interest Points[C] // ICCV. 2003 : 432 – 439.

[20] DOLLáR P, RABAUD V, COTTRELL G, et al. Behavior recognition via sparse spatio-temporal features[C] // In VS-PETS. 2005 : 65 – 72.

[21] PENG X, WANG L, WANG X, et al. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice[J]. CoRR, 2014, abs/1405.4506.

[22] CAI Z, WANG L, QIAO X P Y. Multi-View Super Vector for Action Recognition[C] // CVPR. 2014.

[23] PENG X, ZOU C, QIAO Y, et al. Action Recognition with Stacked Fisher Vectors[C] // ECCV. 2014.

[24] ONEATA D, VERBEEK J J, SCHMID C. Action and Event Recognition with Fisher Vectors on a Compact Feature Set[C] // ICCV. 2013.

[25] CARUANA R. Multitask Learning[J]. Machine Learning, 1997, 28(1) : 41 – 75.

[26] LAWRENCE N D, PLATT J C. Learning to learn with the informative vector machine[C] // ICML. 2004.

[27] BAKKER B, HESKES T. Task Clustering and Gating for Bayesian Multitask Learning[J]. Journal of Machine Learning Research, 2003, 4 : 83 – 99.

[28] EVGENIOU T, PONTIL M. Regularized multi–task learning[C] // SIGKDD. 2004 : 109 – 117.

[29] ARGYRIOU A, EVGENIOU T, PONTIL M. Multi-Task Feature Learning[C] // Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006. 2006 : 41 – 48.

[30] JACOB L, BACH F R, VERT J. Clustered Multi-Task Learning: A Convex Formulation[C] // Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008. 2008 : 745 – 752.

[31] KANG Z, GRAUMAN K, SHA F. Learning with Whom to Share in Multi-task Feature Learning[C] // ICML. 2011 : 521 – 528.

[32] GONG P, YE J, ZHANG C. Robust multi-task feature learning[C] // SIGKDD. 2012 : 895 – 903.

[33] JAYARAMAN D, SHA F, GRAUMAN K. Decorrelating Semantic Visual Attributes by Resisting the Urge to Share[C] // CVPR. 2014 : 1629 – 1636.

[34] KUMAR A, III H D. Learning Task Grouping and Overlap in Multi-task Learning[C] // ICML.

2012.

[35] ZHOU Q, WANG G, JIA K, et al. Learning to Share Latent Tasks for Action Recognition[C] // ICCV. 2013.

[36] KIM S, XING E P. Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity[C] // ICML. 2010.

[37] CORTES C, VAPNIK V. Support-Vector Networks[J]. Mach. Learn., 1995, 20(3) : 273 – 297.

[38] GUYON I, VAPNIK V, BOSER B E, et al. Structural Risk Minimization for Character Recognition[C] // NIPS. 1991 : 471 – 479.

[39] FREY B J, DUECK D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315 : 972 – 976.

[40] HOU R, ZAMIR A R, SUKTHANKAR R, et al. DaMN - Discriminative and Mutually Nearest: Exploiting Pairwise Category Proximity for Video Action Recognition[C] // ECCV. 2014.

[41] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C] // ICCV. 2011.

[42] WANG H, ULLAH M M, KL?SER A, et al. Evaluation of Local Spatio-temporal Features for Action Recognition.[C] // BMVC. 2009.

[43] JÉGOU H, CHUM O. Negative Evidences and Co-occurences in Image Retrieval: The Benefit of PCA and Whitening[C] // ECCV. 2012 : 774 – 787.

[44] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[C] // NIPS. 2014.

[45] SADANAND S, CORSO J J. Action bank: A high-level representation of activity in video[C] // ICCV. 2012.

[46] YANG X, TIAN Y. Action Recognition Using Super Sparse Coding Vector with Spatio-temporal Awareness[C] // ECCV. 2014.

[47] PERRONNIN F, SÁNCHEZ J, MENSINK T. Improving the Fisher Kernel for Large-Scale Image Classification[C] // ECCV. 2010.

[48] PERRONNIN F, DANCE C R. Fisher Kernels on Visual Vocabularies for Image Categorization[C] // 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. 2007.

[49] ZHANG Y, WU J, CAI J. Compact Representation for Image Classification: To Choose or to Compress?[C] // CVPR. 2014.

[50] K. R K, MUBARAK S. Recognizing 50 human action categories of web videos.[J]. Mach. Vis. Appl., 2013, 24(5) : 971 – 981.

[51] ARGYRIOU A, EVGENIOU T, PONTIL M. Convex multi-task feature learning[J]. Machine

Learning, 2008, 73(3) : 243 – 272.

[52] LIU J, JI S, YE J. Multi-Task Feature Learning Via Efficient l2,1-Norm Minimization[J]. CoRR, 2012, abs/1205.2631.

[53] LAZEBNIK S, SCHMID C, PONCE J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories[C] // CVPR. 2006 : 2169 – 2178.

[54] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[C] // NIPS. 2014 : 568 – 576.

致谢

不知不觉在又迎来了在老校区度过的第三个冬天，校园内的景色依旧，而我的心境却再也不是二年半前刚入校的时候了。这二年半的时间中，我无论是在哪一方面的收获都非常的多。而这都要多亏了所有关心我、爱护我和帮助过我的老师和朋友们。

首先，我要特别感谢我的导师邓成教授。他在学术上毫无保留的给我传授他多年的积累和经验，在生活上他教会了我很多做人做事的道理。他教会我做事情一定要尽自己最大的努力做到极致，对科研要抱着最大的热忱坚持到最后。相信我即使我工作之后，我也会继承邓老师教我的对生活和真理的态度，时刻谨记优秀是一种习惯。

感谢实验室的高新波教授，他的谦谦君子风范和严谨的治学态度都给我的留下了深刻的印象，这都会在我今后的人生道路上激励和指引着我。

感谢实验室的李洁老师、韩冰老师、田春娜老师、王秀美老师、王颖老师、张建龙老师、宁贝佳老师、王斌老师、路文老师、牛振兴老师在学术上对我的无私的帮助。感谢您们为我营造和提供良好的科研环境和科研资源。

感谢杨艳华博士，她无论是在科研和生活上对我的真诚的建议都让我获益颇多。感谢吕宗庭师兄、王嘉龙师兄、徐艳红师姐、谢芳师姐、彭海燕师姐、王东旭师兄、邓慧茹师姐、许洁师姐、朱楠师兄在科研和生活上对于我的帮助和关心。感谢和我一起学习科研的唐旭、叶宋杭、吴文军、杨天平、郑睿姣、丁利杰、张相南、杨二昆等同学，这两年半的时光中，我们一起进步，一起成长。感谢刘诣涵、李泽宇、李超、冯英旺、薛雨萌、王浩、陈兆佳、李昭、樊馨霞等师弟师妹。

感谢中国科学院深圳技术研究院的乔宇博士在我实习阶段给我的无私的帮助和建议。感谢三个月中一起学习的杜文斌、王喆、者雪飞、贺盼、张飞云、徐霄、朱细妹、杜书泽、高永强、王浩等同学，他们对科研的热爱和付出给我留下了深刻的印象，与他们的交流令我开阔了眼界和思维。

感谢我的父母，他们是最无私的人，他们是我可以求学的坚实后盾，他们在精神上对我的支持帮助我度过了一个个的难关。感谢他们，未来的路我们还一起走。

最后，衷心感谢各位老师对本论文提出的批评指正!