分 类 号_____ 学号 D201077433

学校代码 10487 密级_____

華中科技大學

博 士 学 位 论 文

物体识别中的形状建模和弱监督学习

学位申请人： 王兴刚

学 科 专 业： 通信与信息系统

指 导 教 师： 刘文予教授

答 辩 日 期： 2014年11月8日

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Engineering

Shape Modeling and Weakly Supervised Learning in Object Recognition

Ph.D. Candidate	:	Xinggang Wang
Major	:	Communication and Information System
Supervisor	:	Prof. Wenyu Liu

Huazhong University of Science & Technology

Wuhan 430074, P. R. China

November, 2014

独创性声明

本人声明所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除文中已标明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，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

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保密□，在 ____ 年解密后适用本授权书。

　　不保密 □。

（请在以上方框内打“√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘　要

人类的物体识别能力是十分神奇的，人能够快速的在复杂场景中，不同光照条件下，识别出数千种类型的物体，不论物体展现出各种不同的姿势，甚至是被部分遮挡。至今，计算机视觉的物体识别算法仍然不能够同人类视觉的感知水平相提并论。在计算机视觉中，物体识别是最核心的问题，从底层的图像特征到高层的视觉应用基本上都是围绕着物体识别展开。物体识别被认为是智能机器人的必备能力，是如今网络上海量视频影像数据检索和挖掘的重要工具，同时也在辅助驾驶，智能安防，人机交互，医学图像分析等领域发挥着极为重要的作用。

本文针对物体识别中的形状建模和弱监督学习这两个具体问题展开了研究，就如何基于形状特征给出更好的物体表达和利用弱监督学习来发掘图像中的语义，克服物体识别中的各种困难，取得更好的物体识别精度，提出了一系列新颖的物体识别方法：

1) 提出了一种基于轮廓片段和词袋模型的物体模型——轮廓片段包。该模型给出了一种新颖的形状表示方法，使得在进行形状的匹配、分类、检索过程中，不再需要寻找形状间点与点之间的对应关系，可以直接在向量空间中得到形状之间的相似度（或者距离），从而快速的完成基于形状的物体识别。另外，轮廓片段包方法中采用了多尺度的形状局部特征，利用区分型学习选择最有区分性的特征和空间配置来完成形状的识别。在众多的形状识别的标准测试集上，该模型取得了目前最高的识别精度。
2) 提出了一种可用于自然图像中物体识别的形状模型——扇形形状模型。该模型基于形状特征和物体部件构建，有效的缩小了形状模型与自然图像中物体识别的语义鸿沟。在进行物体识别时，该模型能够很大程度的对抗物体的形变，容忍物体边缘图像出现破损，并抑制嘈杂背景对物体识别的干扰，快速的估计物体尺寸，结合图像中的纹理特征，取得优异的物体识别精度。
3) 提出了一种在弱监督情况下基于低秩优化的学习物体模型的新算法。在只给定图像标记而不给定物体位置的弱监督的情形下，创新性的将学习物体模型的问题形式化为一个低秩优化问题。该问题是一个凸优化问题，本文采用基于交替方向乘子法可以快速的得到最优解，从而学习到物体的模型并同时定位物体。该方法在物体发现、多示例学习等应用上取得了一系列的成功。
4) 提出了一种新颖的多示例学习方法来学习区分型图像码本并将其用于图像表示和图像分类。不同于传统的多示例学习中只有正负两类示例，本文方法将正示例按照其特征自适应的划分到不同的子类别中，形成一个多类的多示例学习问题。在只给定图像标记而不给定物体位置的弱监督的情况下，有效的

发掘出图像中有区分性的模式，构建简洁高效的图像表示。该方法在场景图像的识别中取得了当时业界最好的结果，并在识别速度上取得了大幅度的提升。

本文是以物体识别为主线，着重于物体的形状建模和弱监督学习，并在这两个关键点上取得了一些突破性成果，已经得到了大量的实验证明。本文所提出的理论、模型及算法对于其它计算机视觉、机器学习方法及应用也有指导意义。

关键词： 物体识别，物体检测，图像分类，物体发现，形状分析，弱监督学习，多示例学习，低秩优化

Abstract

The object recognition ability of human is very amazing. People are able to quickly recognize thousands of kinds of object in clustered scene despite illumination variation, pose variation and occlusion. Nowadays, the ability of computer in object recognition is still not comparable to human. In computer vision, object recognition is a core problem. From low-level image descriptor to high-level vision application, they are surrounding object recognition. Object recognition is considered as an indispensable ability for intelligent robot, and is an important tool to understand big video/image data in internet. Meanwhile, object recognition plays an important role in video surveillance, automatic driving, human-computer interaction and medical image analysis.

This paper aims to study two concrete problems on shape modeling and weakly supervised learning in the context of object recognition, which are how to build better object model based on shape feature and how to use weakly supervised learning to discover visual semantic in image in order to obtain accurate object recognition. Several novel methods have been proposed in this paper:

1) Fan Shape Model, a shape-based and part-based object model, can be used for object recognition in natural image using shape feature. This model bridges the semantic gap between shape model and object recognition in natural image. It is able to tolerant substantial shape deformation, is robust to broken edges, can obtain very impressive detection results even when the edge quality is bad, can fast infer object scale which could also be used by other object detection systems, and can easily combine both shape and texture descriptors for accurate object detection.
2) Bag of Contour Fragment model, based on local shape descriptor and bag of word framework, gives a novel shape representation which is a single vector. The vector representation can be used for shape matching, classification, and retrieval without explicitly finding the correspondence between points on different shapes. Besides, in this model, multiple scale shape descriptors have been used. Furthermore, the most discriminative local features and the spatial configuration are selected via a learning method. It can obtain the state-of-the-art performance on most shape recognition benchmarks.
3) A novel weakly supervised subspace learning method based low-rank optimization. The method is applied to solve this problem of common object discovery. The proposed low-rank optimization formulation can learn object model and localize

object at the same time. The formulation leads to a convex optimization problem, which can be efficiently solved by alternating direction method of multipliers. The method has been successfully applied in object discovery and multiple instance learning.

4) A novel multiple instance learning method for codebook learning to image representation and image classification. Unlike the situation in traditional multiple instance learning where there are only positive instances and negative instances. In this paper, positive instances are divided into different sub-classes based on their features. Thus, the problem becomes a multiple class multiple instance learning problem. The method has been applied to learn discriminative patterns in images to build effective and efficient image representations. In the application of scene image classification, the proposed method achieves the state-of-the-art performance with high classification speed.

The paper focuses on shape modeling and weakly supervised learning in object recognition, and achieves some progresses on the two research topics. The theories, models and algorithms proposed in this paper can also be adopted in other applications in computer vision and machine learning.

Key words: Object Recognition, Object Detection, Image Classification, Object Discovery, Shape Analysis, Weakly Supervised Learning, Multiple Instance Learning, Low-Rank Optimization

目　录

符号对照表

ADMM	交替方向乘子法（Alternating Direction Method of Multipliers）
AP	平均精度（Average Precision）
BoW	词袋（Bag of Word）
BoCF	轮廓片段包（Bag of Contour Fragment）
CNN	卷积神经网络（Convolutional Neural Network）
DCE	离散轮廓演化（Discrete Contour Evolution）
DL	深度学习（Deep Learning）
DPM	可变形部件模型（Deformable Part Model）
EM	期望最大化（Expectation Maximization）
FSM	扇形形状模型（Fan Shape Model）
FPPI	每图平均虚警率（False Positive Per Image）
gPB	全局概率边缘（global Probability Boundary）
HOG	梯度方向直方图（Histogram of Oriented Histogram）
IDSC	内部距离形状上下文（Inner Distance Shape Context）
ILSVRC	ImageNet大规模视觉识别竞赛（ImageNet Large Scale Visual Recognition Competition）
k-means	K均值算法（k-means）
KNN	K近邻（K Nearest Neighbor）
LBP	局部二值化模式（Local Binary Pattern）
LLC	具有局部约束的线性编码（Local-constrained Linear Coding）
LDA	隐式狄利克雷分配（Latent Dirichlet Allocation）
LLE	局部线性嵌入（Local Linear Embedding）
MIL	多示例学习（Multiple Instance Learning）
MMDL	最大间隔多示例码本学习（Max-margin Multi-instance Dictionary Learning）
NMS	非最大抑制（Non Maximum Suppression）
PR	精度-召回率（Precision-Recall）
RBF	径向基函数（Radial Basis Function）
RPCA	鲁棒主成分分析（Robust PCA）

RANSAC	随机采样一致（Random Sample Consensus）
SC	形状上下文（Shape Context）
SIFT	尺度不变的特征变换（Scale Invariant Feature Transformation）
SVD	特征值分解（Singular Value Decomposition）
SVM	支持向量机（Support Vector Machine）
TILT	变换不变的低秩纹理（Transform Invariant Low-Rank Textures）
VOC	视觉物体分类（Visual Object Classes）
WSL	弱监督学习（Weakly Supervised Learning）

1 绪论

1.1 选题背景

视觉是人和动物感知外部环境的主要途径。动物需要在自然环境中通过定位和识别物体来判别天敌、躲避障碍物等。人类可以分辨超过30000个类别的物体，并在数百微秒之内定位物体[①]。历经三十余年的研究，计算机视觉已经取得了突飞猛进的成果，尤其在于物体识别方面。但尽管如此，目前计算机视觉的性能同人类视觉的性能仍不能相提并论。作为计算机视觉中的核心问题，物体识别还是计算机视觉领域的研究重点。

伴随着计算机视觉的不断发展，计算机视觉离人们的日常生活越来越接近。试想：当一位大意的行人突然出现你驾驶的汽车面前，汽车是否可以自动提醒你刹车从而避免车祸? 当你拍照时，相机是否可以自动找到拍摄画面中的人脸或景物从而自动对焦? 当你看到一件喜欢的衣服，是否可以通过拍摄照片，上传电商网站自动搜索到同样款式的衣服并购买? 计算机是否能够从医学图像（如CT图像）中自动的发现有病变的器官组织来辅助医生诊断疾病? 这些只是众多需要从图像中进行物体识别的应用中的四个例子。如今，在智能安防、互联网、机器人、人机交互、医学图像等众多领域对于物体识别的需求越来越大，解决好计算技术视觉中的物体识别问题有助于推动国民经济的发展。

1.1.1 物体识别中的具体任务

计算机视觉中需要解决的任务可以分为底层任务（low-level task）和高层任务（high-level task）。底层任务包括三维重建、图像匹配、图像分割、图像边缘提取等，这些任务一般都不需要去理解图像中语义，直接利用图像中的像素信息就可以完成。高层任务的目的是去理解图像中的语义（通常也被称为高层语义）。例如，理解图像中的物体类别或人物身份，即物体识别；理解图像中人的动作，即动作识别（包括手势识别、姿态识别）；理解视频中人的行为，即行为识别；跟踪视频中的特定物体，即视频跟踪；更多的一些高层语义的识别还可以是去理解图像中人物的心情，理解蒙娜丽莎微笑的涵义等。

上述识别任务当中，物体识别是计算机视觉里所有高层任务中最基础且最重要的。如图1-1所示，它的范畴十分广泛，包含很多的具体的任务。这些不同的具体的任务都得到了研究人员的充分研究。具体来讲，按照物体识别的对象级别的不同，物体识别可以划分为：

① 数据来源于加利福尼亚大学伯克利分校计算机视觉组主页：http://www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/

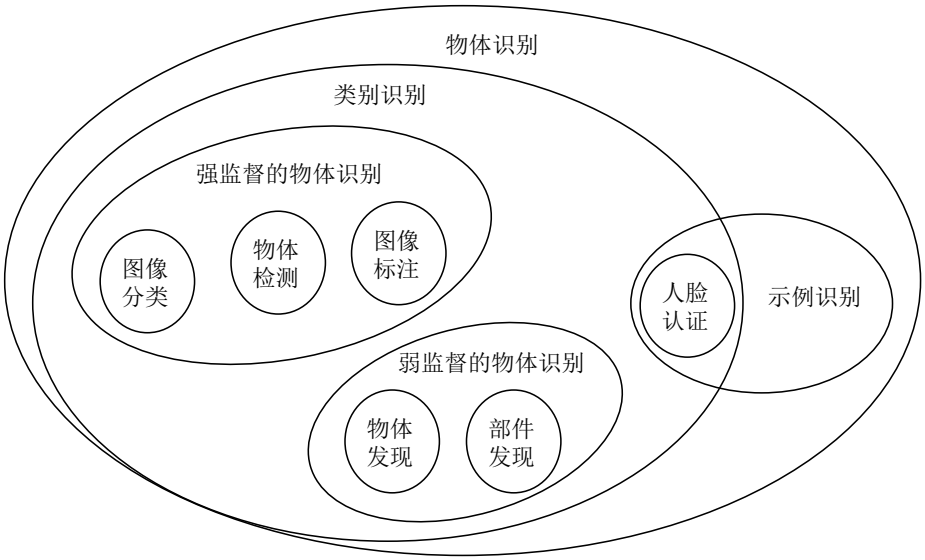


图 1-1 物体识别中包含的各种任务

1) 示例识别，其目的是为了判断不同图像中的物体是否对应于同一个特定的示例。如同一只鞋子、同一个人的脸。传统的示例识别通常采用无监督的图像特征点匹配来完成。
2) 类别识别，其目的是为了判定图像中物体的类别。类别的定义是灵活的，可以根据实际的应用环境进行调整。实际应用中，类别的定义具有层次性，例如，在人脸检测中，人脸区别其它的物体可以作为一个类别；在区分不同人种时，黄种人和白种人的人脸属于不同的类别；在区别人的身份，两个不同身份的黄种人的人脸属于不同的类别。

随着类别识别中类别地不断细化，类别识别从一定程度上可以包含一些示例识别，例如，上面例子中判定图片中的人脸是否属于同一个人这个任务，即人脸认证，既可以被理解为示例识别，也可以被理解为类别识别。但是，对于视觉上极为相似的不同物体，如两个同一款的手机，很难去定义它们是否属于不同类别。往往，在这种情况，可以采用视频中的帧间信息来区别不同示例。类别识别应用更为广泛，也更为重要。

根据物体识别的粒度，从整幅图像，到部分图像区域，再到图像中的每一个像素，物体识别可以划分为：

1) 图像分类，其目的是判定图像中是否含有特定类别的物体。宽泛来讲，场景图像的分类也属于物体识别，尽管它并不关注于特定类别的物体。
2) 物体检测，其目的不但需要判定给定图像中是否含有特定类别的物体，而且还需要给出该类别物体的位置。通常位置用物体的包围盒（Bounding Box）表示，也有一些情况下位置用物体的外轮廓（Contour）更精确的表示。

3) 图像标注，其任务需要对于图像的每一个像素给出一个物体类别。由于应用中关注的物体类别有限，通常，不关注的物体类别所对应的像素均被归类到背景类。

从图像分类，物体检测到图像标注，物体识别的粒度从大到小变化。在完成小粒度的物体识别任务后，可以轻松的得到大粒度的物体识别任务的结果。另外，大粒度物体识别任务的正确结果也可辅助将小粒度物体识别任务完成的更好。

图像分类、物体检测和图像标注的另外一方面的区别是它们所需要的监督信息（Supervision）不同。物体识别系统一般都包含训练（Training）步骤和测试（Testing）步骤。从机器学习的角度，按照训练物体识别模型所给定监督信息和期望在测试图像上输出的信息，物体识别可以分为：

1) 强监督的物体识别，在强监督的物体识别中，训练阶段给定的标注信息和期望在测试图像上得到的输出信息的形式是一样的。例如，在给定图片级标注的情况下进行图像分类；在给定包围盒级标注的情况下进行物体检测；以及在给定像素级标注的情况下的图像标注。
2) 弱监督的物体识别，在弱监督情况下，训练阶段给定的标注信息的格式要弱于期望在测试图像上得到的输出信息的格式。例如，在给定图片级标注的情况下，要求定位图像中的物体位置，这个任务被称为物体发现；在给定包围盒级标注的情况下，要求输出物体部件的位置，这个任务被称为部件发现。

强监督的物体识别是传统的物体识别问题。在如今大数据时代，强监督的物体识别虽然制约于需要人工的标记的高成本和低效率，但由于众包（crowd-sourcing）地迅速发展，有监督的物体识别可以通过Amazon Mechanical Turk等众包平台获得数百万记的标记信息，这使得强监督的物体识别得到了迅速的发展。弱监督的物体识别虽然不能够准确的输出预期的结果，如准确的定位物体或准确的找到物体的部件，但是弱监督的物体识别往往能够帮助更好的完成强监督的物体识别。而且，在大数据时代，弱监督有天然的优势——弱监督的物体识别所需要的标记信息少，从而容易获得更多的训练数据。

在上述的各种物体识别的任务中，物体检测最为重要的。在自然场景中，往往物体检测是图像分类的先决条件——只有定位了物体才能够准确地判决物体的类别。物体检测亦可以辅助图像标注，提高图像标注的速度和精度。物体发现可以视作为一种在弱监督情况下的物体检测。另外，部件发现往往被用于学习物体检测模型，并作为物体检测方法的一部分出现。因此，物体检测是物体识别的核心任务。

1.1.2 物体识别中面临的困难和发展历程

物体识别是计算机视觉中最有挑战性的任务。计算机可以利用多个视角拍摄的图像重建出场景的3D形状，但它却无法判别场景中物体的类别，即便是以一个两岁

小孩的水平来衡量也是如此。认知科学的研究者尚且不能探索出人类识别物体的机理奥秘。从计算机视觉的角度来讲，物体识别面临诸多困难：

1) 形变：现实环境中的物体分为刚性物体和非刚性物体。刚性物体往往不存在形变，例如椅子、汽车、手机等人工制造的物体。除了刚性物体之外，更多的物体是非刚性的，如人体的姿势变化会造成不同的人体形状的显著不同，人的表情变化会使人脸的外观有显著的差异。通常，物体的计算模型（Computational Model）可以将一个非刚性物体分解成为多个刚性部件，并用类似于弹簧的可变形的链接来连接这些刚性部件。
2) 视角变化：在现实环境中三维的物体被投影到二维图像时，由于相机拍摄视角的不同，得到的二维物体图像会有显著的不同。小的拍摄视角的变化，可以近似的当作形变来处理。但是大的视角变化，会彻底的改变物体模型的结构。例如，从正面来拍摄汽车和从侧面来拍摄汽车得到的物体在外观和形状上都有显著的不同。如何采用的统一的模型去表示不同视角下的物体尚没有定论，目前多采用混合模型（Mixture Model）来处理视角变化带来的问题。
3) 遮挡：遮挡会导致物体的部件不可见，物体模型在图像上不能完全匹配。如何在物体只有部分可见的情况下识别出物体，并推断出物体不可见部件的位置？对于这些问题，目前通常将物体分解为不同部件，构建基于部件的物体模型。
4) 嘈杂背景：现实环境中物体所在的背景可能会十分的复杂。例如，狮子会处在杂草从中，鸟会出现在树叶之间等。这些干扰信息存在丰富的纹理可能会被识别成物体而造成错误的识别。
5) 成像条件不同：同一个物体，在不同的成像条件上会呈现出不同的外观。过度曝光、光照过暗、反光、相机运动模糊等情况都会对物体识别带来不同的困难。
6) 尺度变化：物体所在的数字图像中，物体的尺寸可以大至数万像素，小至不足100像素。通过图像金字塔技术可以实现多尺度的物体识别，但是当物体过小，对应的图像中包含的信息有限，即使采用超分辨技术也不可能将物体扩大到一个合适的尺寸。在这种情况下，如何实现物体识别仍然有待研究。
7) 类内差异：同一类物体，不同个体，在不同时间上往往会存在巨大的差异。例如，人穿不同的衣服会有不同的外观，同一个人的人脸的照片在幼儿、青年和老年会有巨大的差异。

以上7点是物体识别中的主要困难，图1-2对应给出了一些具体示例图像。在实际情况中，上述困难中的多个困难往往会同时出现在同一幅图片或同一段视频中。面临着这些困难，物体识别方面的研究人员设计了一系列有效的物体识别方法。本文将按照时间顺序叙述主要的程碑式的物体识别方法。

图 1-2 物体识别中的困难，包括：形变、视角变化、遮挡、嘈杂背景、成像条件的不同、尺度变化和类内差异。

特征点匹配方法 计算机视觉中引用次数最高的方法是由Lowe提出的SIFT（Scale Invariant Feature Transformation）方法[1]②。SIFT是最早在图像中提取关键点上的特征，并通过匹配这些特征来实现物体识别的方法。SIFT利用高斯差分（Difference of Gaussian）方法找到图像中对于尺度、形变、光照等干扰因素稳定的特征点及其对应的区域，然后在这些区域上计算梯度直方图特征。这种方法在很大程度上克服了形变、尺度变化、成像条件不同等困难，得到了稳定物体识别性能，特别是在示例识别方面。同时，该方法也推动了三维重建、图像检索等应用的发展。SIFT方法开启了特征点匹配系列方法的先河，至今在这个研究线路上仍然有大量的后续研究工作。Mikolajczyk等人在文献[2,3]中，对于如何在图像中提取的稳定的关键点以及如何更好地去描述图像中的局部区域中作了详尽的调研。SIFT在鲁棒性和准确性上已经非常成功，但它的计算速度略慢，后续有一系列工作研究如何加快局部特征的计算速度，如SURF[4],ORB[5]等，以及适合在智能终端上使用的LDB[6]方法。另外，Belongie等人提出的形状上下文（shape context）方法[7]是一种有效的形状特征，在形状匹配上取得了巨大的成功。基于特征点匹配的物体识别的主要优点是：对于大部分示例识别精度高。它的主要缺陷是：（1）对于每幅图像，特征点的数量在数百到一两千的样子，每一幅图像都被表示为一个特征点的集合。在进行物体识别的时候，需要计算集合同集合之间的距离，如Hausdorff距离、Hungarian距离等。因此，

② 目前根据Google Scholar上统计的SIFT方法的引用次数（包括期刊版本和会议版本）超过33000次

在物体识别的速度很慢，特别当待识别的图像在达到上万的级别。（2）特征点匹配方法的泛化能力差，不能够克服类内差异的问题，很难将属于同一个类别的不同示例匹配在一起。

特征袋模型 针对特征点匹配方法中速度慢、泛化能力差的问题，特征袋方法开始流行，并在物体识别中发挥作用。特征袋方法的原型是自然语言处理领域用于文档分析的词袋（BoW，Bag of Word）方法，最先由Sivic和Zisserman引入到计算机视觉领域中用来物体识别，具体为物体检索[8]，后由Csurka等人将其与分类器联系在一起用来做图像分类[9]。特征袋模型的思路是根据一个特征点的码本，统计一幅图中所有的特征点在码本中各个不同码字上出现的频率。特征点的码本通常采用聚类算法得到，如k-means方法，后来也发展出各种不同的码本学习方法，如层次码本[10]、区分型码本[11]、稀疏型码本[12]。特征袋模型中最原始的频率统计是将每一个特征分配到码本中的一个码字，这个过程称为矢量量化（Vector Quantization）。之后，矢量量化开始往特征编码（Feature Coding）发展，特征编码的目的是寻找码本中的一个子集，并调整它们的系数，使得这个子集和这些系数能够更精确地表示当前特征。特征编码的里程碑工作是由Yang等人提出的采用稀疏编码（Sparse Coding）来进行特征编码的方法[13]，后续的研究工作包括局部约束的线性编码[14]、Fisher核编码[15]、径向核编码[16]、VLAD[17]等方法。

原始的特征袋模型当中，特征之间的空间关系完全被忽略，这限制了特征袋模型的物体识别精度。Lazebnik等人把由Grauman和Darrell提出的金字塔匹配核（Pyramid Match Kernel）[18]引入到了特征袋模型当中，提出了空间金字塔（Spatial Pyramid）方法[19]。空间金字塔方法可以灵活的应用图像中特征之间的空间关系，显著地提高了特征袋模型的物体识别精度。空间金字塔的后续工作中包含本人提出的特征上下文方法[16]，该方法利用了极坐标系来刻画图像特征间的空间关系，在众多数据集上取得了比空间金字塔方法优的图像分类和物体检测精度。

特征袋模型已经被广泛地应用于物体检测、图像分类、图像检索。在这三种应用中，特征袋模型所需要的码本大小也不仅相同，一般物体检测小于图像分类，图像分类小于图像检索。因为物体检测和图像分类只关注特定类别的物体，且物体检测关注的类别比图像分类关注的类别少，同时物体检测和图像分类都可以采用分类器进行特征选择，然而图像检索是无监督的，不能够利用分类器来选择一些关键的特征，只能是采用一个大的码本来隐性的完成不同图像的特征集之间的匹配。特征袋模型已经被广泛地应用于物体检测、图像分类、图像检索。在这三种应用中，特征袋模型所需要的码本大小也不仅相同，一般物体检测小于图像分类，图像分类小于图像检索。因为物体检测和图像分类只关注特定类别的物体，且物体检测关注的类别比图像分类关注的类别少，同时物体检测和图像分类都可以采用分类器进行特征选择，然而图像检索是无监督的，不能够利用分类器来选择一些关键的特征，只

能是采用一个大的码本来隐性地完成不同图像的特征集之间的匹配。特征袋模型已经被广泛地应用于物体检测、图像分类、图像检索。在这三种应用中，特征袋模型所需要的码本大小也不仅相同，一般物体检测小于图像分类，图像分类小于图像检索。因为物体检测和图像分类只关注特定类别的物体，且物体检测关注的类别比图像分类关注的类别少，同时物体检测和图像分类都可以采用分类器进行特征选择，然而图像检索是无监督的，不能够利用分类器来选择一些关键的特征，只能是采用一个大的码本来隐性地完成不同图像的特征集之间的匹配。特征袋模型能够快速完成物体识别的原因是它给图像的特征集提供了一个向量表示，计算向量之间的速度要远远快于计算集合之间的距离。正因为特征袋方法能够快速的计算图像之间的距离，特征袋模型能够采用更多的图像样本作为训练，从而得到高精度的物体识别性能。

特征袋模型中的图像特征往往采用前面提到的SIFT、SURF等。最初研究人员倾向于采用稀疏的尺度不变的特征，然而随着图像数据的增多、识别问题的规模变大，研究人员发现稠密的多尺度的特征能够更好的描述图像，在采用特征袋模型的情况下得到更高的精度[19]。尽管如此，特征袋模型的表达能力还是有限的，其主要制约因素在于：（1）特征袋模型中的特征是手工设计的，手工设计的特征不免会存在丢失图像中有效的可用于物体识别的信息；（2）特征袋模型过于简单，在提取完图像中的特征之后，模型包含一个特征编码层和一个判别层（如，采用SVM、最近邻等分类器），缺乏灵活性，表达能力不强。

统计学习模型 特征袋模型在图像局部特征的基础上利用图像码本统计计算给出了一个简洁的图像表示。后续出现的统计学习模型侧重于在给定冗余的图像特征基础上，利用机器学习中的统计学习算法，选择有效的图像表示方式。其中最具代表性的工作有两个：Dalal和Triggs提出的人体检测方法[20]，我们可以称之为DT方法；Viola和Jones的人脸检测算法[21]，一般简称为VJ方法。在DT方法中，通过计算图像中各个像素的梯度方向，统计出梯度方向的直方图，并采用不同的归一化方式给出了一个信息含量丰富的图像表示，即HOG（Histogram of Oriented Gradients）特征。然后采用线性SVM来选择HOG中有助于识别物体的维度。在VJ方法中，利用图像中不同区域的像素差值作为特征，即Haar特征，Haar特征可以通过积分图像快速地计算。在采用VJ方法进行人脸检测中，对于每一个人脸检测的窗口，采用160000个Haar特征，这160000个Haar特征被当作AdaBoost[22]中的160000个弱分类器，最后由AdaBoost选择出200最有区分性的Haar特征来快速地完成人脸的检测。VJ方法充分体现了统计学习模型这一类方法的精髓，后续出现了大量相关的研究工作在计算机视觉的应用中取得了众多的成果。统计学习模型的强项在于刚性物体的检测，其弱点在于它不能够很好的处理物体的形变。

基于局部的物体模型 Felzenszwalb等人提出的可变形部件模型（DPM，Deformable Part Model）在非刚性物体的检测性能上取得了性能上的飞跃[23]。DPM方法将非刚性物体分解成为不同的刚性部件，对于每一个部件采用HOG+线性SVM的统计学习模型表示，并采用一个星形结构来表征不同部件的空间关系，从而能够对抗物体的形变。国内中科院的研究组基于DPM提出了一系列新的改进，并取得了PASCAL VOC物体检测的第一名[24]。针对大规模的多类别物体检测，Google公司的研究人员结合DPM和Hash技术在单台PC机上实现了十万个物体类别的快速物体检测[25]。就目前来看，DPM系列的方法的主要问题在于它的学习能力有限，不能利用海量的图像信息来学习出比HOG、LBP（Local Binary Pattern）等手工设计的特征更加有效的图像表达。

深度学习方法 目前最有效的图像识别方法是深度学习方法。深度学习方法基于神经网络，由Hinton和Krizhevsky等人[26]在ILSVRC[27]大规模的图像识别上应用并取得了突破性的进展③。在传统的卷积神经网络（CNN，Convolutional Neural Network）[28]基础上，深度学习方法采用了更大更深的网络结构（Hinton和Krizhevsky等人的网络中含有6千万个参数和65万个神经元）以及Dropout、Pre-training等技术，利用GPU计算实现了网络的训练。深度学习技术的突破不仅仅表现在图像分类上，还表现在自然场景文字识别[29]、语音识别[30]等。

模型驱动和数据驱动，二者哪一个更好？这是一直以来计算机视觉（以及其它很多的计算机科学中的研究领域）中面临的一个问题。深度学习系列方法在经典的优美的神经网络模型下，采用数据驱动的形式，取得了一系列令人振奋的研究进展。从这个角度上来讲，深度学习的研究回答了这个问题，那就是：模型驱动和数据驱动都十分重要，只有在二者完美结合的时候才能够真正解决实际中问题。

1.1.3 目前物体识别的问题和本文的提出的解决思路

以上本文对于物体识别的内容和方法进行了归纳和总结，探讨了各类物体识别方法的优势和不足。目前，从结果来上来看，深度学习方法在大规模的图像分类上都取得了十分令人振奋的结果。最新的ILSVRC 2014的图像分类结果已经由Google团队采用改进的深度学习技术将错误率降低到6%，这是一个非常低的错误率。尽管如此，物体识别问题离完美解决还有很远的路途。正如最新的在计算机视觉数据集上的研究工作Microsoft COCO项目[31]表明：ILSVRC中采用的大规模数据集ImageNet中的图像主要是图标图像（Iconic Image），自然图像中的物体识别要比图标图像中的物体识别困难很多。相对于图标图像中的物体识别，自然图像中的物体识别首先需要解决物体检测问题。Girshick等人[32]尝试利用一些通用的物体检测

③ ILSVRC 2012比赛中，总共有来自于1000个类别的1000000图像。Hinton等人的深度学习方法将错误率从之前最好的0.26172到0.15315。

器（Generic Object Detector）在图像中生成物体的候选区域，然后采用深度学习方法去分类这些候选区域。这个方法在一些数据集上取得了良好的物体检测结果，但该方法还是不能够绕过物体检测过程，尽管他们采用的是通用的物体检测器。针对如今的物体检测任务，其核心问题是如下两个：

- 高效的物体模型，物体模型对应于物体表达（Object Representation）。其中高效的含义包括两层：（1）对于形变、类内差异等上节叙述的7点物体识别中的变化具有鲁棒性，在统一的物体模型下尽可能多的去克服识别困难；（2）可快速计算的特性，物体检测需要采用滑动窗口（Sliding Window）技术去扫描图像中的各个可能存在物体的区域，如果物体模型不可以被高效计算，将会使得对应的物体检测算法不可用。
- 有效的学习方法，人工设定的物体模型参数往往是不可靠的，物体模型的参数需要通过以一种数据驱动的形式从训练数据中学习得到。学习物体模型的目标方程基本上都是非凸的，甚至是组合优化等复杂形式，另外，训练图像的数量巨大，导致计算难度大。因此需要有效的方法去调整物体模型中的参数，从而尽可能多地吻合训练图像中物体的存在的形式，并保持在测试图像上的泛化能力。

针对上述两个核心问题，本文提出以形状建模和弱监督学习两个思路分别来探索解决方案。

高效的物体模型需要高效的图像特征，而且形状特征是一个十分优异的图像特征，在人识别物体的过程中形状也起着关键的作用[33]。在计算机视觉中，形状特征具有对物体上纹理变化、颜色变化、光照变化的鲁棒的特征，另外，形状特征擅长于灵活地表征物体的形变[7,34,35]。因此本文的研究中将以形状特征为基石来建模物体。

基于形状特征来构建物体模型一直以来都得到了整个计算机视觉领域的高度关注，之前的研究方法在构建高效的物体模型方向做出了积极的探索，但仍然存在一些局限性：

- 一些经典的形状模型，如形状上下文[7]、内部距离形状上下文[34]、形状树[36]等方法可以较好的建模二值形状，但是它们无法跨越二值形状与自然图像之间的语义鸿沟（Semantic Gap），均不能直接用于自然图像中的物体识别。
- 另外，目前这些形状模型不具备可快速计算特性。在采用它们进行物体识别中，往往依赖于集合之间的匹配[7,37,38]，甚至是图匹配[34,36]。集合匹配和图匹配都需要计算局部特征的对应关系，非常耗时。本文的研究将文档分析中的词袋模型引入到形状建模当中，对形状给出了一个向量形式的简洁表示，满足了形状模型的在识别中的可快速计算特性。

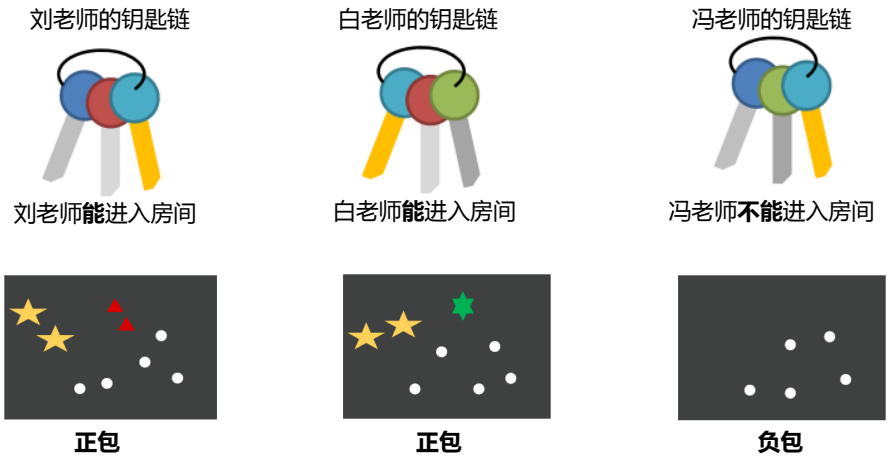


图 1-3 多示例学习示意图。图中三串钥匙中有两串可以打开房间，另一串不可以，我们可以轻易的推测出能够推测出哪一类钥匙是可以打开房间的，其它的钥匙则不能。在多示例学习当中，一串钥匙对应于一个包，其中的一根钥匙对应于包中的一个示例。能打开房间的钥匙对应于正示例，反之则对应于负示例。正包中肯定包含正示例，负包中全为负示例。

- 新兴的一些方法尝试将利用形状特征在自然图像中进行物体识别[39–44]。但这些方法均只关注于采用各种方法将训练集中的形状（或物体轮廓）同测试图像中的边缘匹配起来，缺少一个显性的物体模型，一个既能够表达物体的结构，也能表示物体外观的模型。缺少一个显性的物体模型的问题在于：当训练样本集中的形状变多，这些自底向上的方法会有更多的轮廓模板需要匹配，导致物体识别速度变慢。借鉴经典的图案结构（Pictorial Structure）[45,46]的原理，本文给出一个结构化的基于形状和部件的物体模型。

计算机视觉和机器学习始终是两个密切相关的研究领域。机器学习中的很多方法都在计算机视觉中的诸多应用中发挥着巨大作用，如条件随机场（CRF，Conditional Random Filed）[47]，支持向量机（SVM，Support Vector Machine）[48]等；计算机视觉的研究也推动了很多机器学习方法的发展，加深人们对于机器学习方法的理解，如实时的人脸检测算法中的AdaBoost[21]，大规模图像识别中的深度卷积神经网络[26,28]。在物体识别中，学习技术显得尤为重要，因为我们需要将人的高层语义通过学习融入到识别过程中。目前有大量的强监督的学习方法被用于目标识别，在上节介绍的“统计学习模型”和“深度学习方法”中给出了一些具体的例子。

尽管如此，采用弱监督的学习方法进行物体识别并没有得到深入的研究。Viola等人[49]将多示例学习（MIL，Multiple Instance Learning）和AdaBoost结合，在弱监督的情况下完成人脸检测。但该工作只是提出了弱监督情况下物体识别这个想法，并没有给出具有说服力的实验验证。后续，Babenko等人[50]将多示例学习和AdaBoost结合这一思路成功地应用在视频跟踪中。目前，将弱监督学习应用于物体识别中最成功的方法是上节介绍的DPM方法。DPM方法中提出了具有隐变量

的SVM（Latent SVM）算法，其本质也属于多示例学习算法。图1-3中通过一个简单的例子解释了多示例学习的本质。正如这个例子中多示例学习可以找到能打开房间的钥匙，多示例学习也可以挖掘出图像中对于物体识别有帮助的关键信息。因此，本文在研究物体识别中的弱监督学习问题时均基于多示例学习框架。然而，现有的多示例学习算法依旧局限于Stuart等人[51]提出将分类器和多示例限制迭代优化的策略上，并未就物体识别中具体的情形做相应的改进，具体表现为：

- 未考虑图像中不同物体部件的不同外观。现有的多示例学习方法均只区分正示例和负示例，是一个二类学习问题。然而在物体识别中，物体所对应的正示例在特征上是差异性很大，将它们强制的归为一类会限制物体识别的性能。本文的研究将对应于物体的正示例按照其外观特征的不同，自适应地划分到不同的子类别中，从而得到一个多类的多示例学习问题。
- 目前用于物体识别的多示例学习的方法均在套用现有的分类器，如SVM、Boosting、Random Forest[52]等，其目标方程的求解均涉及到组合优化问题，该问题严重非凸。现有方法均无法保证全局最优解，这限制了弱监督情况下物体识别的结果。本文的研究舍弃了现有的分类器，将物体模型与多示例学习放置在统一的数学框架下，提出一个凸优化的目标方程，并寻求它的全局最优解，从而更好地完成弱监督情况下的物体识别。

综上，本文就物体识别中的建模和学习问题，针对性地从形状特征、部件模型和多示例学习等方向展开了研究。

1.2 本文的研究内容和贡献

本文的研究内容在于**探索物体识别中两个核心问题：（1）形状建模问题——如何构建稳健高效的基于形状和部件的物体识别模型，和（2）弱监督学习问题——如何在弱监督的情况下有效的通过机器学习方法去自动地发现图像中的共同物体或部件。**针对了这两个核心问题，本文提出了若干创新性算法，在图像分类、物体检测和物体发现三个应用上取得了优异的实验结果。图1-4中对这三个任务做了一个直观的阐释。图1-5中展示了本文的主要研究内容，下面将就形状建模和弱监督学习这两个方面分别展开介绍。

形状建模问题 形状是人类进行物体识别时采用的重要特征，而且好的物体模型是物体识别关键，因此利用形状特征来构建物体识别模型是计算机视觉中的重要问题。这个问题的难点在：

1) 自然图像中形状特征的不稳定性。在自然图像中提取物体的形状（边缘）并不完美，物体的边缘往往存在缺失，同时除了物体之外，还有很多的属于背景的噪声边缘。

图 1-4 本文所研究的物体识别中的图像分类、物体检测、物体发现任务。图中左侧显示各个任务在训练阶段给定的标注情况，图中右侧对应各个任务在测试阶段期望输出的结果。图中实线表示标记，虚线表示期望的输出。

2) 物体形状间的局部相似性。物体在形变、遮挡、视角变化等情况下，物体的形状并非全局相似而是局部相似。

针对这两个难点，本文的研究思路是将物体分解为不同的部件，采用基于部件的物体表示方法。这个思路对上述两个难点给出了统一的解决途径：当物体边缘缺失的情况下，采用图像中物体的可见部件来推断缺失部件的位置，从而完成精确的物体识别；当同类别物体之间只有局部相似的情况下，选择出相似的部件进行物体识别。在这种思路下，本文研究（1）如何采用形状特征去描述物体部件，（2）如何从图像中提取物体部件，（3）如何描述不同的物体部件之间的空间关系，以及（4）如何快速的确定图像中的物体相对于物体模型的尺度等问题。针对物体检测和图像分类两个具体的任务，本文在形状建模方面提出了两个具体的方法：轮廓片段模型和扇形形状模型。

弱监督学习问题 弱监督学习原本属于机器学习研究的范畴，鉴于计算机视觉同机器学习的紧密联系，本文在物体识别这个范畴之下探索弱监督学习的问题。具体来讲，在物体识别中本文研究在之给定图像级标记情况下去挖掘同类别图像中的共同物体，即物体发现。由于监督信息弱，这是一个十分困难的问题。假设我们以物体的包围盒来表示物体所在的位置，对于500×600像素大小的图像，我们可以从中提取数百万个的不同宽高比的包围盒。假如给定1000张包含共同物体的图像（每幅图像中只包含一个共同物体），那么我们的任务就是数十亿的样本中找到1000个对于共同物体的样本。这是一个组合爆炸的问题，传统的机器学习方法根本无法解决这

个问题。因此本文就如何快速的解决这一组合爆炸问题，提出了两种思路，一种是基于低秩优化的思路，假设所有的共同物体存在于同一个低秩的子空间中；一种是基于区分型学习的问题，通过训练分类器将共同物体与背景区分开来。以上两种不同的思路可以得到了不同的目标方程，本文研究如何求解对应的凸优化问题和半凸优化（semi-convex optimzation）问题。另外，本文将机器学习中的多示例学习同本文中的研究方法结合在一起，分析其中的内在关系，研究如何更好地解决多示例学习问题。最后，本文将物体发现同强监督的物体识别任务结合起来，研究如何利用弱监督情况下的学习到的物体信息来更好的完成强监督的物体识别任务，提出最大化间隔的多示例码本学习方法用于图像分类任务。

本文研究内容之间的联系 本文中研究内容之间的联系可以用图1-5中的箭头连线表示。本文中的两个核心研究内容均可以用于解决物体识别中的图像分类、物体检测和物体发现这三个问题。物体的形状模型结合分类器或通过匹配可以直接完成图像分类和物体检测任务；弱监督学习方法的输出便是图像中的共同物体，从而完成物体发现任务；另外，物体的形状模型可以作为物体的表示在物体发现中使用，从而更好的挖掘出图像中在形状上具有（局部）相似性的物体，并克服同类物体上不同纹理的干扰；弱监督学习方法可以发现图像中的语义，作为“属性”（attribute[53]）来用于图像分类和物体检测。对于这两个核心研究内容，本文分别提供了两个研究方法，并对应一系列研究问题，如基于部件的物体表示、低秩优化、多示例学习等。在图1-5中，实线联系表示在本文中已经完成研究，虚线联系表示对应研究工作在本文中尚未处理，将作为本文的后续研究。

本文的贡献 计算机视觉和机器学习均是目前计算机科学中热门的研究领域。本文以物体识别为目的，将物体的形状表示与弱监督学习紧密结合，提出了一系列新颖的算法，解决了图像分类、物体检测、物体发现等任务中的关键问题。下面将按照本文中的四个主要研究方法来介绍本文的贡献。

1) 在轮廓片段包模型中，本文首次将自然图像中的先进的特征编码和空间金字塔方法引入到形状表示中；本文提出了多尺度的形状部件特征，并自动选择合适的尺度进行最有效的物体识别。在绝大部分的形状识别测试集上，获得目前最高的分类和检索准确率。另外，轮廓片段包模型对于形状给出了一个向量表示，具有识别速度快的优点。
2) 在扇形形状模型中，本文提出了一个新颖的基于部件和轮廓的物体模型。该模型具有（1）可自动利用物体轮廓学习而无需人工标注部件，（2）可克服物体边缘缺失，（3）可快速确定图像中物体尺寸，（4）可灵活的对抗物体的形变等优点。在标准的基于形状的物体检测数据集上取得了优异的物体检测性能。

3) 在基于低秩优化的子空间发现方法中，本文首次将物体的低秩表示（Low-Rank Representation）用于物体发现任务，仅采用一个简洁的凸优化方程就可以完成物体发现任务。提出了单类的多示例学习问题，并给予解决思路。在自然图像中的物体发现和多示例学习两个任务中取得了良好的结果。
4) 在最大化间隔的多示例学习中，本文创新性的提出多类的多示例学习方法，并通过最大化间隔的形式化求解这一个问题。本文将最大化间隔的多示例学习用于构建图像表示的字典学习中，提供了一个区分性强的简洁的图像表示，取得了优异的图像分类性能。

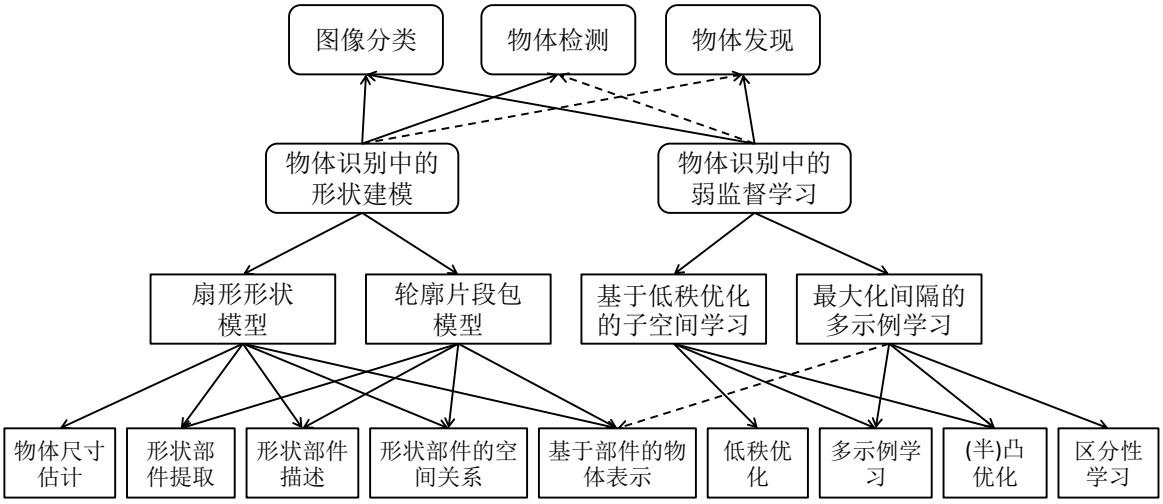


图 1-5 本文的主要研究内容

1.3 本文章节安排

本文章节安排如下：

第1章介绍首先研究背景，包括物体识别的研究内容，面临的困难，以及发展历程，从而引出本文的选题思路。之后介绍本文的研究内容、研究意义和研究的贡献。

第2章提出扇形形状模型，首先给出基础的形状表示，即基于射线的形状表示，并研究采用动态规划方法来完成形状匹配；接下来定义物体模型，利用形状匹配的结果来估计物体模型的参数；然后在得到物体模型中，研究如何快速地确定物体尺寸，并完成精确的物体检测。另外，此章中还介绍了如何采用低秩优化来克服形状识别中的旋转不变性的问题。最后展开实验验证，包括形状分类、形状聚类、物体检测等任务。

第3章提出轮廓片段包模型，首先提出多尺度的轮廓片段特征，包括该特征的提取方法和描述方法。其次给出如何将该特征编码，采用空间金字塔进行汇聚，得到

简洁的形状表达。接下来介绍如何采用线性SVM分类器进行形状的分类。最后测试轮廓片段包方法的形状分类及形状检索性能，讨论该方法的鲁棒性及该方法中参数对于识别性能的影响。

第4章提出基于低秩优化的子空间发现方法，首先介绍子空间学习和物体发现的研究现状，然后给出基于低秩优化的子空间发现问题的形式化方法，接下来给出基于ADMM的快速求解方法，最后采用仿真和物体发现的实验验证方法的有效性。

第5章提出最大化间隔的多示例学习方法，首先介绍相关研究工作，然后给出最大化间隔的多示例学习的形式化方法，给出由该形式化带来的优化问题，提出一种迭代优化的解决方案。在给出最大化间隔的多示例学习之后，将其应用到图像的码本学习中，给出一个具有区分性的图像表示。最后展开实验验证本章方法在图像分类上的有效性和合理性。

第6章总结全文，讨论本文的研究工作的新颖性、实用性、启发性，并展望本文的后续研究工作。

2 轮廓片段包模型

形状的表示是计算机视觉中的一个基础问题。之前的形状表示方法主要关注于设计一些对于形状旋转、尺度和形变不变的底层描述符，本章研究内容关注于中层次的形状建模，提出了一个新颖的简洁的形状表示——轮廓片段包（BoCF，Bag of Contour Fragment）。在轮廓片段包方法中，形状被按照内在结构分解成不同的轮廓片段，并被编码成形状码（shape code），最后通过一种考虑形状空间结构的汇聚（Pooling）方法得到一个形状的向量表示。轮廓片段包方法适用于形状分类、形状检索，在众多标准的形状测试集上取得了目前最好的精度，且具有速度快的优点。

2.1 研究现状

形状可以直观的体现图像中的语义，即使很小的孩子都可以从简单的线条中识别出各种物体。另外，在自然图像中形状特征对于光照、物体颜色、纹理变化鲁棒。这些优点使得形状在物体识别中得到了广泛的应用。

经过多年来形状领域的研究，大量的不同的形状描述符被提出用来形状的匹配和识别。基于形状区域的形状描述子有Zernike矩[54]和通用的Fourier描述子[55]。基于形状轮廓的形状描述子有曲率尺度空间方法[56]（CSS，Curvature Scale Space）、多尺度凹凸方法[57]（MCC，Multi-scale Convexity Concavity）、三角形区域表达[58]（TAR，Triangle Area Representation）、形状树[36]（shape-tree）、轮廓灵活度方法[59]（contour flexibility）、形状上下文[7]（SC，shape context）和内部距离形状上下文[34]（IDSC，Inner-Distance Shape Context）等。在本章中提出的轮廓片段包方法中，上面介绍的基于轮廓的形状描述子中均可以作为其中的底层描述。为了不失一般性，本章中采用其中最为常用的一个简单的形状描述子，形状上下文，作为底层描述。

基于这些形状描述子，形状识别技术得到了迅速的发展。Sun和Super[60]提出了一个采用轮廓片段作为特征和采用Bayesian分类器进行分类的形状识别框架。Bai等人[61]结合轮廓片段和骨架片段在Sun和Super的工作基础上取得更好的结果。Daliri和Torre[62,63]将形状轮廓上的点转换成一个符号表示（symbol representation），然后采用编辑距离（edit distance）来度量两个形状之间的距离。Wang等人[64]提出要采用骨架联合树（tree-union）[65]表示每个形状类别的原型，并采用各个不同类别的联合树来进行形状识别。Edem和Tari[66]同样采用了骨架的树模型来表示每个形状类别的模型，单个形状到不同的骨架树之间的编辑距离被作为特征，形状识别由线性SVM完成。更多不同的形状识别方法[67–70] 在这里不再赘述。

在之前的研究工作中，轮廓片段（contour fragment）特征被证实为一个非常有效的形状描述特征。但之前的研究工作在轮廓片段上的工作只是简单的将形状看成

一个轮廓片段的集合，形状识别时需要计算集合到集合之间的距离，这样的计算过程费时同时难以选择出具有区分性的轮廓片段。本章将研究在轮廓片段特征的基础上进一步学习，发掘具有代表性的轮廓片段，对于整个形状抽象出一个简洁的向量化的表示。本章提出的方法称作为轮廓片段包方法，是一个两层的特征学习框架。在第一层中，采用LLC方法轮廓片段被编码成形状码，LLC方法最先由Wang等人[14]提出用于编码自然图像中的SIFT特征。其它的一些编码方法如Fisher核[15] 等方法也可以用在本章方法当中。之所以选择LLC作为本章中的编码方法主要原因在于LLC方法具有速度快精度高的优点。在第二层中，采用空间金字塔匹配[18,19].（SPM，Spatial Pyramid Matching）在考虑轮廓片段间的空间关系基础上汇聚形状码得到简洁的形状表示。最近的深度学习（DL，Deep Learning）系列方法产生了巨大的影响，在大规模的图像识别领域获得了巨大的成功[26]。不同于本章中提出的轮廓片段包方法，深度学习方法拥有更多的层数。在形状建模方面，Eslami等人[71]提出深度的形状伯尔兹曼机器（SBM，Shape Boltzmann Machine）直接从形状的像素入手学习一种概率分布来建模形状。在形状补全（shape completion）任务上取得了好的结果，但该方法不能应对形状上显著的旋转、形变，不适合用来完成分类、检索等识别任务。目前，仍没有比较适合形状识别的深度学习方法。图2-1展示了轮廓片段包方法构建形状表示的流程，本章接下来的内容将逐步介绍其中的各个部分。

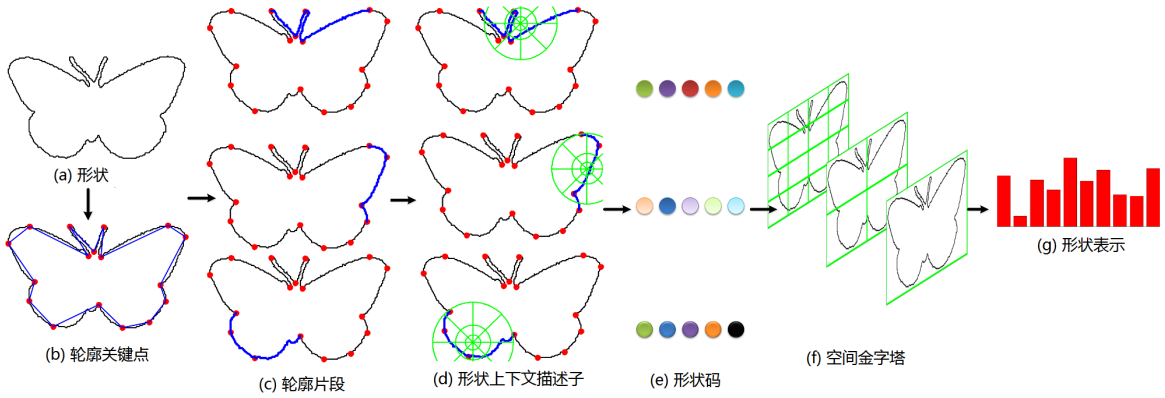


图 2-1 采用轮廓片段包进行形状表示的流程。（a）中展示了一个形状；（b）中显示了采用DCE 方法在形状中找到的关键点；（c）中采用蓝色显示了一些轮廓片段；（d）显示了采用形状上下文特征[7]来描述每一个轮廓片段；（e）表示形状片段被编码成为形状码；（f）中采用1×1，2×2，和4×4的空间金字塔来进行最大汇聚（max-pooling）；（g）中显示了最终的形状的向量表示。

2.2 轮廓片段特征

轮廓片段是从物体的完整形状上提取，通过多尺度的提取轮廓片段，轮廓片段可以包含形状的局部和全局信息。在之前的形状识别的研究中，轮廓片段已经

被验证是一种十分有效的形状特征[60,61]。本章方法中采用轮廓片段作为基础的形状特征，然后学习一个形状的码本和建立形状表示。将一个物体的轮廓分解为轮廓片段有不同的方法，包括对轮廓的均匀采样和基于曲率的采样[60]等。本章方法使用的是一种叫做离散轮廓演化（DCE，Discrete Contour Evolution）[72]的技术将轮廓分割成有意义的片段。DCE方法利用形状的全局相似性得到形状轮廓上的关键点，相对于轮廓上的局部极值点，DCE方法得到的轮廓关键点更稳鲁棒。定义$\{S(t) = (x(t), y(t)), t \in [0, 1]\}$为一个形状的$S$的外部轮廓④，其中$x(t)$和$y(t)$分别表示轮廓上点的横纵坐标。首先使用DCE方法将$S$简化为多边形，多边形的顶点表示为

$$\overrightarrow{u} = (u_1, \ldots, u_T),$$

对应于形状轮廓的**关键点**，其中T表示的是关键点的个数，$\forall i \in [1, \ldots, T], u_i \in [0, 1]$。关键点的个数是通过给点的阈值$\tau$自动计算得到的，而不需要人工设定。阈值$\tau$控制的是简化得到的多边形与原始形状之间的相似度，$\tau$越小，简化得到的多边形与原始形状之间越相像，多边形的顶点数目也越多。图2-1(b)给出了从轮廓S上通过DCE提取关键点的过程。

在得到形状中的关键点之后，形状可以根据这些关键点分解为不同的轮廓片段。定义一个物体轮廓S的**轮廓片段**集合为$\mathcal{C}(S)$，其中每一个轮廓片段由一对关键点之间轮廓构成，因此定义c_{ij}为关键点u_i和u_j之间的轮廓片段，即

$$c_{ij} = \{S(t), t \in [u_i, u_j]\} \tag{2-1}$$

根据以上定义之下我们可以将轮廓片段集合表示为

$$\mathcal{C}(S) = \{c_{ij}, i \neq j, i, j \in [1, \ldots, T]\} \tag{2-2}$$

需要的注意的是，关键点u_i和u_j不一定是相邻的。同时还有

$$S = c_{ij} \cup c_{ji}, \tag{2-3}$$

c_{ij}和c_{ji}互为补集，它们共同构成形状S。可以发现本章中提取的任意关键点对之间的轮廓片段是多尺度的，由轮廓片段组成的集合$\mathcal{C}(S)$提供了形状S的丰富信息。概括来讲，这种信息可以概括为为短范围，中范围和长范围。图2-2显示了这种多尺度信息。因此，轮廓片段与图像中的局部纹理描述符，如SIFT，HOG，LBP等，完全不同。后者描述的对象是图像的局部小块，不如轮廓片段特征具有多尺度性。本章的后续部分将会说明如何表述轮廓片段和如何选择信息量大的轮廓片段进行形状识别。

④ 当物体形状存在内部轮廓时，可以采用外部轮廓和各个不同的内部轮廓分开，分别采用DCE算法。

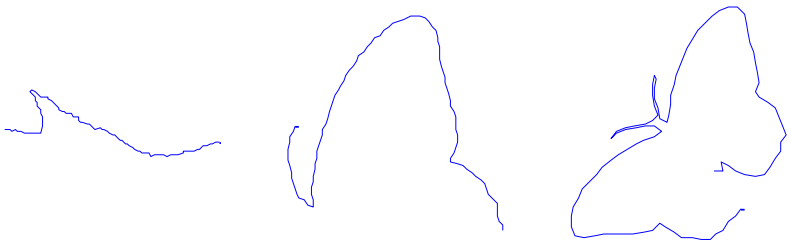


图 2-2 一个形状中选取的三个轮廓片段，分别对应于形状中的短范围、中范围和长范围的信息。

对每一段轮廓片段c_{ij}，本章中使用经典的形状上下文描述符[7]来表述，对于同一个轮廓片段，可以采用多个形状上下文特征拼接成为一个向量$\mathbf{x}_{ij} \in \mathbf{R}^{d\times 1}$，其中$d$是多个形状上下文特征的总维度，即$c_{ij}$的特征维度。图2-3直观的描述了基于形状上下文特征计算轮廓片段特征的过程。对于轮廓片段的特征$\mathbf{x}_{ij}$，其具体的计算过程描述如下：首先在c_{ij}上从u_i到u_j等距的采样5个点。对于不同长度的c_{ij}，采样点的数目固定，因为如此才能得到固定长度的特征表示。采样点数目的可以调整，5是一个经验值，太多的采样点会导致特征计算、编码比较慢，太少的采样点会使特征的表达能力变弱。在这5个采样点上分别计算形状上下文特征，之后将这5个形状上下文特征拼接起来形成一个向量作为轮廓片段c_{ij}的特征。在本章中，$\mathbf{x}_{ij}$是形状S的基本的描述子。

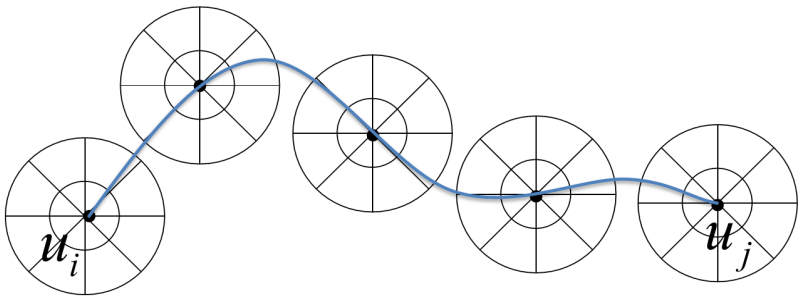


图 2-3 采用形状上下文特征来描述轮廓片段c_{ij}的示例。图中的圆圈仅示意计算形状上下文特征的位置，并不表示实际计算形状上下文特征的区域。

2.3 轮廓片段的编码

在特征包的图像表示框架内，特征编码是一个十分关键的步骤，它对于最终物体识别的性能影响很大。特征编码的作用在于将原始的特征映射到一个高维空间中，在高维空间中，可以将编码后的特征进行汇聚得到简洁的形状表达。这个映射过程是通过一个形状码本（Shape Codebook）$\mathbf{B}$实现的，映射后的空间用符号$\mathcal{B}$表示。在本章中，一个轮廓片段特征表示为$\mathbf{x}_{ij}$，轮廓片段编码后的特征为$\mathbf{w}_{ij}$。在基

于特征包的图像表示研究中，很多码本学习的方法被提出了，其中包括无监督的方法[14]和有监督的方法[73,74]。码本学习并不是本章中研究的重点，因此采用一种简单而且稳定的K-means[75] 算法作为本章码本学习的方法。具体的学习码本的过程描述如下：首先从数据集中所有的用于训练的轮廓片段特征中随机选择出一部分，然后对随机选择出来的部分用K-means算法进行聚类。聚类中心就作为是形状的码本$\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_M] \in \mathbf{R}^{d \times M}$。其中码本的每一列就是一个聚类中心，可以被认为是形状空间中的一个基。一般来讲，根据本人经验，如果希望学习得到一个大小为M的码本，则需要随机选取$M \times 500$个轮廓片段特征以保证k-means得到的形状码本具有代表性。

得到形状码本之后，轮廓片段包方法便可以基于形状码本进行特征编码，这个编码过程也可以被称作是量化过程。在所有的编码方法中，最简单的方法且最传统的方法为向量量化方法（VQ，Vector Quantization）[19]，VQ将形状特征$\mathbf{x}_{ij}$只分配给形状码本$\mathbf{B}$中与其最近的聚类中心。这种方式虽然很高效，但是造成的量化误差很大。局部约束线性编码（LLC，Local-constraint Linear Coding）[14]是最近流行的一个特征编码方法，它具有快速而且有效的特点。LLC的编码方式是受到了发表在《科学》杂志上的局部线性嵌入（LLE，Local Linear Embedding）[76]理论的启发。为了在形状码本生成的特征空间$\mathcal{B}$上表示$\mathbf{x}_{ij}$，LLC使用$\mathbf{x}_{ij}$在形状码本中的k个最近邻建立一个局部坐标系。定义$\mathbf{x}_{ij}$的k 个最近邻为$\mathbf{B}_{\pi_{ij}} \in \mathbf{R}^{d \times k}$，其中$\pi_{ij} = \{\pi_{ij}^1 \ldots \pi_{ij}^k\}$是码本$\mathbf{B}$中的$k$个最近邻的索引集合，则$\mathbf{B}_{\pi_{ij}}$是包含$\mathbf{B}$的$\pi_{ij}^1 \ldots \pi_{ij}^k$列向量的矩阵。依照LLE的假设，$\mathbf{x}_{ij}$和其最近邻位于或者接近一个在流形中的局部线性区域，则$\mathbf{x}_{ij}$和$\mathbf{B}_{\pi_{ij}}$的局部几何关系可以通过线性系数描述。这些线性系数可以通过码本$\mathbf{B}_{\pi_{ij}}$对$\mathbf{x}_{ij}$进行重建得到。具体计算方法为：系数$\mathbf{w}_{\pi_{ij}} \in \mathbf{R}^{k \times 1}$可以通过解决如下的最小化问题得到

$$\min_{\mathbf{w}_{\pi_{ij}}} \|\mathbf{x}_{ij} - \mathbf{B}_{\pi_{ij}} \mathbf{w}_{\pi_{ij}}\|^2 \qquad \text{s.t.} \quad \mathbf{1}^T \mathbf{w}_{\pi_{ij}} = 1, \tag{2-4}$$

其中权重向量$\mathbf{w}_{\pi_{ij}}$描述了在$\mathbf{x}_{ij}$重建中局部基的贡献，权重向量的和限制为1。公式(2-4)中的最小化问题是一个时间复杂度为$\mathcal{O}(k^2)$的小规模最小二乘问题。在实验中，设定k值为5。最终定义$\mathbf{x}_{ij}$的编码结果为$\mathbf{w}_{ij} \in \mathbf{R}^{d \times 1}$。$\mathbf{w}_{ij}$中$\pi_{ij}^1 \ldots \pi_{ij}^k$ 位置的值等于$\mathbf{w}_{\pi_{ij}}$，其它的位置的值设为0。

2.4 基于空间金字塔的特征汇聚

本节介绍在形状码$\mathbf{w}_{ij}$的基础上，融合特征之间的空间关系建立一个紧凑的形状表达。在建模轮廓片段的空间关系的时候，轮廓片段方法中利用了空间金字塔匹配（SPM，Spatial Pyramid Matching）[19]。建立形状表达的过程如下：首先，需要将形状分成不同的区域，如图2-1(f)所示，形状按照1×1, 2×2, 4×4 的方式被

分成总数为21个区域。然后，对每个区域$Region_r, r \in [1,\ldots,21]$进行最大化汇聚（max-pooling），即对特征的每一维取区域中所有特征的最大值。定义$\mathbf{w}^z$为在形状位置z上的编码后的轮廓片段，轮廓的位置由其轮廓中间点的位置决定。最大汇聚用公式表达如下：

$$\mathbf{f}(S,r) = \max(\mathbf{w}^z | z \in Region_r), \tag{2-5}$$

其中最大函数返回一个区域$Region_r$的特征向量$\mathbf{f}(S,r)$。特征向量的长度与$\mathbf{w}_{ij}$是一致的。对每一个码本中编码，将区域中所有的形状编码的最大值作为形状表达，这个过程被称为最大汇聚。最大汇聚过程在视觉皮层（V1）[77]的生理研究中得到了确认。它的正确性在很多的图像分类算法中都得了经验性的证实[13–15]。最大汇聚和线性分类器结合能有比较好的结果，不需要采用计算复杂的非线性分类器。最后，形状S的表示$\mathbf{f}(S)$是所有区域的特征向量的连接。

$$\mathbf{f}(S) = [\mathbf{f}(S,1)^T, \ldots, \mathbf{f}(S,21)^T]^T. \tag{2-6}$$

从上面式子可以的看出，对于每个形状总共有21个区域，每个区域的特征维度为M，最后特征的维度是$\mathbf{f}(S)$ is $21 \times M$。

SPM方法在不同层次上由粗到细地表达了轮廓片段的空间信息。训练一个分类器可以自动判断其在粗尺度（1×1区域）或者是细尺度（2×2和4×4区域）进行形状的识别。特别需要指出的是，如果训练形状能对齐的很好，每一个小的网格区域就包含类似的轮廓片段。这种情况下，分类器应该选择在细尺度上的特征。另一方面，如果训练图像在各个方向上有旋转，那么每个网格区域包含的轮廓片段就不同。但是粗的尺度包含了所有的轮廓片段，那么在这种情况下，分类器在粗尺度上起的作用更大。因此我们可以总结来说SPM是一种十分灵活的描述特征间空间关系的方法。另外，结合一些形状的对齐算法，如本文3.4节中介绍的方法，SPM能够取得更好的结果。为了同其它的方法进行公平的比较，本章的实验中并没有额外的添加形状对齐步骤。

2.5 基于线性SVM的形状分类

在完成特征编码和特征汇聚之后，轮廓片段方法得到的形状表达是一个简单的向量，所以可以直接利用SVM来进行形状分类，而不需要进行复杂的形状的匹配。针对多类的SVM分类问题，本章使用Crammer和Singer提出的多类分类策略[78]。给定一个训练形状$\{\mathbf{f}_i\}$的集合，它的标注为$\{y_i \in [1,\ldots,N]\}$，其中N是形状类别的个数。Crammer和Singer的多类SVM可以直接用来解决如下的优化问题：

$$\min_{\omega_1,\ldots,\omega_N} \quad \sum_{n=1}^{N} \| \omega_n \|^2 + \lambda \sum_i \max(0, 1 + \omega_{r_i}^T \mathbf{f}_i - \omega_{y_i}^T \mathbf{f}_i), \tag{2-7}$$

其中$r_i = \arg\max_{n\in[1,\ldots,N],n\neq y_i}\omega_n^T\mathbf{f}_i$。在公式(2-7)中，左边的部分是正则项，右边的部分是多类问题的hinge损失函数。参数λ控制正则项的相对权重。在实现过程中本章采用Lin等人开发的SVM库LibLinear[79]来求解公式(2-7)。在测试阶段，形状的类别由以下的方程进行预测：

$$\widehat{y} = \arg\max_{n\in[1,\ldots,N]}\omega_n^T\mathbf{f}. \tag{2-8}$$

SVM学习实际上就是一个选择支持向量的过程的，对应在本章的形状识别的具体应用就是选择那些轮廓片段在形状分类中起到重要作用。图2-4给出了选择出来的一些重要的轮廓片段。对于某一类形状，根据$\mathbf{f}_i \cdot \omega_{y_i}$的值由大到小选择排名前20的轮廓片段，可以选择在$\mathbf{f}_i$中贡献最大的前20的轮廓片段。换句话说，上述过程寻找的是在$\mathbf{f}_i$ 中编码值最大的前20个样本。可以从图2-4看出，通过编码（LLC），最大汇聚和SVM分类器选择出来的轮廓片段是有意义的。在前20的轮廓中也有一些不重要的，例如，(c)中的第17和第19。因为它们比较简单能够被很精确的编码，所以编码的值能够很大，但是其对应的ω_{y_i}却很小。

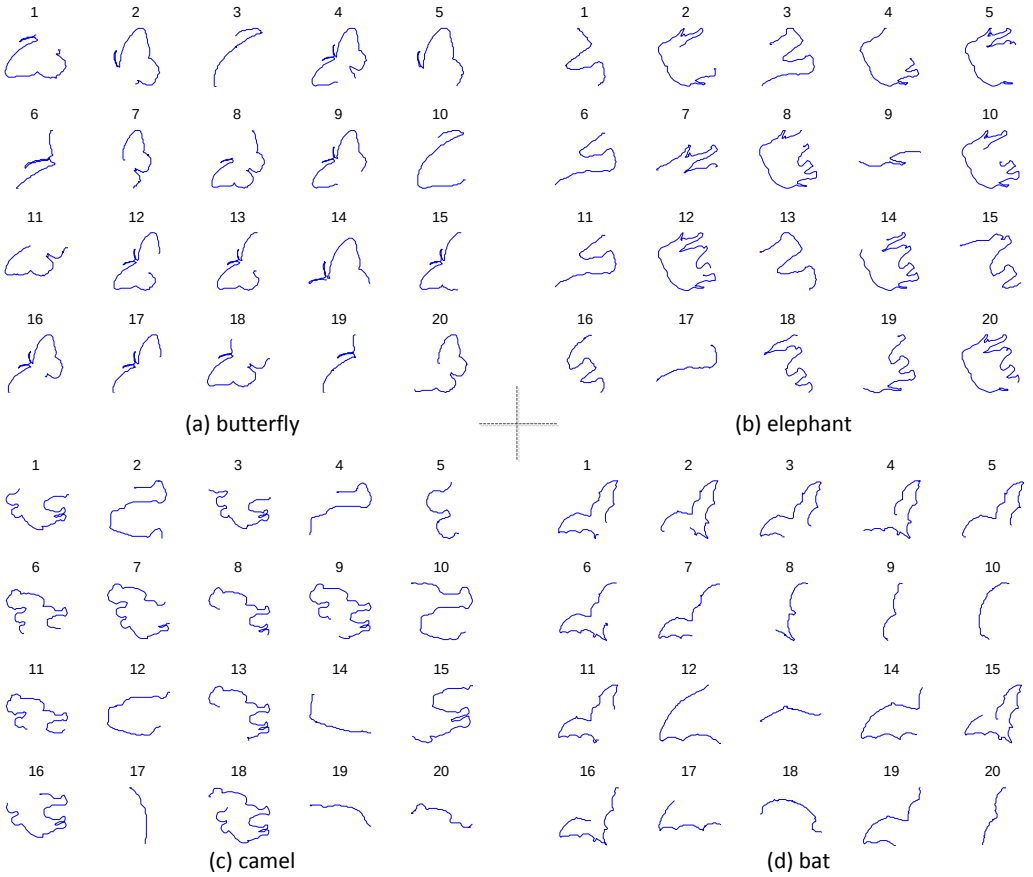


图 2-4 利用本章算法从MPEG-7数据集上选择的对于识别butterfly(a)、elephant(b)、camel(c)、bat(d)四个类别的形状最重要的前20个轮廓片段。

一些更耗时的核函数，例如RBF核和intersection核等，能够进一步提高形状分类的准确度。但是为了更快，我们选择线性核的SVM作为分类器。

2.6 实验

本节给出在不同的形状数据集上测试本章中提出的形状分类算法，并在同样的实验设置下同目前处于领先水平方法进行结果比较。同时，还通过实验来验证本章方法的鲁棒性及其它特性。本章实验中的代码可在如下网址下载：https://bitbucket.org/xinggangw/bcf。

2.6.1 实现细节

提取轮廓片段 本章方法在每个形状中使用DCE算法提取了大约400个轮廓片段；DCE中最大曲率值阈值τ设置为0.5。在计算轮廓片段的形状上下文时，如图2-3中所示，设有5个参考点，同时将形状上下文中的扇形区域数设为60个（角度空间分为10份，半径空间分为6份）。另外，形状中轮廓片段位置同样被记录下来用于利用SPM来编码它们的空间分布信息。

学习形状码本 用标准的k-means聚类方法来进行码本的训练。由于从数据集中提取的轮廓片段的总数目会非常大，如果采用所有的轮廓片段进行码本的训练有很大的时间和空间复杂度。实际中通过采样方法来降低这个复杂度，随机地选取了1000个形状，对于每一个形状，选出300个轮廓片段特征来训练码本。如果未具体指出，聚类中心的数量为1500。另外，实验中会在不同的聚类中心数量下研究本章方法的性能。

编码、汇聚、分类 对于编码方案，在LLC方法中k近邻的数目设置为5。汇聚时，一个形状将会被分成1×1，2×2，4×4不同层次共21 个子区域。最后将表示形状的特征向量进行ℓ_1范数的归一化。对于形状分类，使用一个已有且快速的线性SVM工具包LibLinear[79]。

数据集 在形状分类标准数据集MPEG-7[80]、动物数据集[61]、Swedish Leaf数据集[34]以及ETH-80数据集[81]上测试BoCF方法。在剩余的章节中，将按照不同的数据集给出实验结果和分析。

2.6.2 MPEG-7数据集

MPEG-7数据集在计算机视觉领域被广泛用于形状分析的研究中。它包含来自于70个类别的1400个二值形状，每一个类别有20个存在差异的形状（具体请参考图2-5中一些典型的形状）。将使用以下两种方案来评价形状分类的性能：(1)**半训练**

表 2-1 MPEG-7数据集上的分类准确率比较

算法	分类准确率	
	半训练半测试法	留一测试法
Class segment set[60]	90.9%	97.93%
Contour segments[61]	91.1%	-
Skeleton paths[61]	86.7%	-
ICS[61]	96.6%	-
Polygonal multi-resolution[83]	-	97.57%
String of symbols[82]	-	97.36%
Robust symbolic[62]	-	98.57%
Kernel-edit distance[63]	-	**98.93**%
BoCF	**97.16±0.79**%	**98.93**%

半测试法（Half training），在每一轮中，从每个类别中随机选取10个形状作为训练，将剩下的另一半形状作为测试；此过程重复10 个轮次；给出平均的分类准确率和其标准差。(2)**留一测试法**（Leave one out），对于每一个形状，使用除当前选取作为测试形状之外的所有其它形状作为训练，给出平均分类准确率。

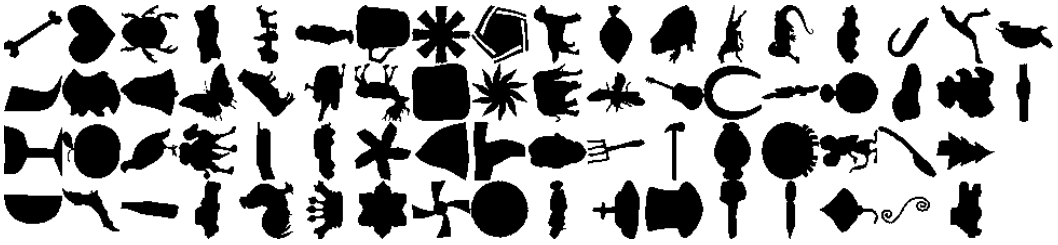


图 2-5 MPEG-7数据集中的一些典型形状。每个类别中选择了一个形状。

表2-1将BoCF与其它形状分类方法在MPEG-7数据集上的分类准确率作了比较。在文献[60,61]中，轮廓片段也被用来进行形状分类。在使用半训练半测试法时，BoCF的准确率优于文献[60,61]中的方法6%。这种明显的性能提升归功于BoCF方法可以通过SVM可以最大化不同形状类别之间的差别，同时对于每一类找到那些信息量大的轮廓片段。然而，在文献[60,61]中，所有的轮廓片段拥有相同的权重。我们不能忽视的一个事实是，BoCF给出了一个向量化的形状表示而且LLC通过精确的编码保留了轮廓片段中的信息，这是可以采用SVM进行区分性学习的前提条件，这充分证实了本章提出的BoCF方法的优点。在文献[62,63,82]中，形状描述基于符号表示法（symbolic representation）。当使用留一测试法时，BoCF与最近研究工作[63]中的性能相同，取得了该数据集上形状分类的最好结果。

2.6.3 动物数据集

动物数据集（Animal Dataset）在文献[61]中提出，它包含了来自于20种动物的2000个形状，如马、兔子、猴子等，每个类别有100张动物形状。在图2-6显示出的数据集里面一些典型形状，其中包含一些比较困难的类别（猫和猴子）和一些比较简单的类别（蜘蛛）。由于相同类别的动物有很大的外观上的差异，因而这些数据集有相当多的类内差异性。对于每一个类别，随机选取的50个形状用来训练，剩下的形状用来测试，进行了10次实验，并统计每一次实验的分类准确率。BoCF的平均准确率在表2-2中给出并与其它的一些优秀的方法进行比较。

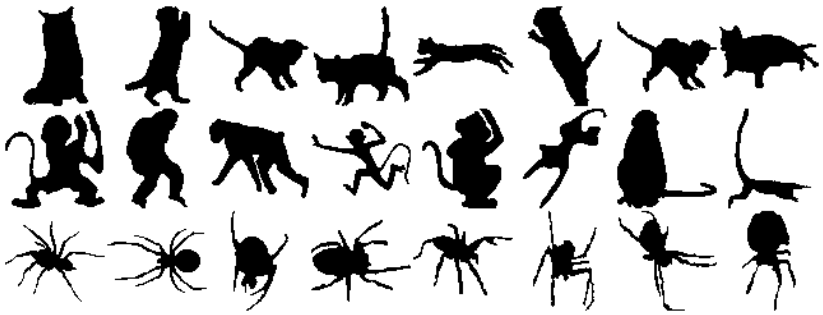


图 2-6 动物数据集中的一些形状。每一行显示来自于同一类别的8个形状，第一行为猫，第二行为猴子，第三行为蜘蛛。

表 2-2 动物数据集上的分类准确率比较

算法	分类准确率
Class segment set[60]	69.7%
IDSC[34]	73.6%
Bag of SIFT[84]	74.9%
Contour segments[61]	71.7%
Skeleton paths[61]	67.9%
ICS[61]	78.4%
BoCF	**83.40±1.30**%

如表2-2中所示，本章提出的BoCF方法得到了83.40%的分类准确率，此准确率明显优于经典形状描述子——IDSC方法[34]和整合轮廓片段和骨架路径的ICS方法。表2-3中对动物数据集中20个类的每一类分别给出了分类准确率。BoCF在猫和猴这两类中极大地提升了准确率，这个说明BoCF可以从动物数据集的高度变形的物体之间抓获类内的部分相似性。基于SIFT的词袋算法(Bag of SIFT)[84]直接使用结构特征作为形状分类，达到了74.9%的分类准确率，比BoCF的准确率低了很多。这个说明

表 2-3 不同算法在动物数据集上每一类单独的形状分类准确率

算法	Bird	Butterfly	Cat	Cow	Crocodile	Deer	Dog	Dolphin	Duck	Elephant
CS[61]	76%	89%	39%	70%	54%	69%	69%	87%	83%	95%
ICS[61]	76%	**93%**	48%	**80%**	66%	79%	75%	**89%**	**89%**	**97%**
BoCF	**87.6%**	92.2%	**73.8%**	77.4%	**76.8%**	**90.4%**	**82.6%**	**89.0%**	87.0%	95.2%
算法	Fish	Fly-bird	Hen	Horse	Leopard	Monkey	Rabbit	Rat	Spider	Tortoise
CS[61]	70%	57%	89%	96%	56%	21%	81%	52%	98%	81%
ICS[61]	74%	65%	94%	**97%**	65%	33%	**87%**	**84%**	**100%**	90%
BoCF	**79.8%**	**72.0%**	**94.2%**	95.4%	**66.4%**	**58.4%**	85.8%	70.6%	99.2%	**93.6%**

本章提出的轮廓片段特征比SIFT更适于形状分类。

2.6.4 瑞典树叶数据集

本节中讨论在瑞典树叶数据集（Swedish Leaf Dataset）[85]上使用BoCF进行树叶图像识别。此数据集来自于Linkoping大学和瑞典自然历史博物馆的一个树叶分类项目，它包含来自于15个不同的瑞典树品种，每个品种有75片单独的树叶。图2-7中展示的是一些典型的树叶形状二值图，其中一些品种对于人眼来讲都很难分辨，例如，第1，3，9，11，15个树品种。按照文献[34]中的设置进行实验：在每个品种中，随机选择25个形状作为训练，剩下的形状作为测试。识别精度的统计方法是：进行10次训练和测试，给出平均分类准确率和标准差。表2-4中给出了BoCF方法与其它基于形状的树叶识别方法进行分类准确率的比较。比较的方法中包括使用像图像矩（moments），面积以及曲率这些简单特征的初级研究工作方法[85]，傅里叶(Fourier)描述子[34]，采用形状上下文和动态规划的匹配方法[34]，采用内部距离形状上下文和动态规划的匹配方法[34]，多尺度矩阵的距离矩阵方法[86]，形态学描述子[87]，鲁棒的符号表示法[62]和形状树方法[36]。在这些方法中，BoCF取得了目前最好的树叶识别精度。

图 2-7 Swedish leaf数据集中的典型形状。其中每一个类别中选择了一个形状。

2.6.5 ETH-80数据集

ETH-80数据集[81]包含有来自于8个类别的80个高分辨率的三维物体（如图2-8所示）。对于每一个物体，包含从不同视角观察的41张彩色图像，因而此数据集共

表 2-4 Swedish leaf数据集上的分类准确率

算法	分类准确率
Moment+Area+Curvature[85]	82%
Fourier[34]	89.6%
SC+DP[34]	88.12%
IDSC+DP[34]	94.13%
MDM[86]	93.60%
IDSC+Morphological strategy[87]	94.80%
Robust symbolic[62]	95.47%
Shape-tree[36]	96.28%
BoCF	**96.56**±**0.67**%

有3280张图像。通过在彩色图像中将物体分割出来，便可以得到三维物体投影到不同平面的形状，这些利用这些投影得到的形状来来分析评价基于形状的物体识别方法。依照之前文献中的习惯性做法[81]，在这个数据集选择留一测试法的交叉验证的测试模式。具体地说，在每一轮将来自79个物体的图像用来训练，剩下一个图像用来测试。在表2-5中给出了BoCF方法的平均准确率和很多其它方法进行了对比。BoCF 达到了91.49% 是目前最好的识别效果。

表 2-5 ETH-80数据集上的分类准确率的比较

算法	分类准确率
Color histogram[81]	64.86%
PCA gray[81]	82.99%
PCA masks[81]	83.41%
SC+DP[81]	86.40%
IDSC+DP[34]	88.11%
IDSC+Morphological strategy[87]	88.04%
Height function[88]	88.72%
Robust symbolic[62]	90.28%
Kernel-edit[63]	91.33%
BoCF	**91.49**%

图 2-8 ETH-80数据集中的80个三维物体。每一行的物体来自于同一个类别。

2.6.6 对噪声的鲁棒性

在上述的实验中，所用的数据集形状都是自然界中的形状，因此形状的轮廓比较平滑。为了测试BoCF 在噪声情况下的性能，我们可以在形状边缘上加进高斯噪声后使用BoCF进行形状分类。本次实验中使用整个MPEG-7 数据集作为没有添加噪声的原始数据。采用如下方式添加噪声：对于形状轮廓上的每一点，包括X轴和Y轴方向，都添加均值为0方差为σ的高斯函数随机生成的噪声。当σ增加时，形状轮廓上就会加上了更多的高斯噪声。图2-9展示了不同程度的高斯噪声下的形状边缘。图2-10中给出了当噪声函数的方差σ的值从0变到1的时，记录下使用半训练半测试法和留一测试法的分类准确率。当σ值从0增到1时，使用半训练半测试法的分类准确率下降了大约4%，这说明BoCF方法对于噪声具有鲁棒性。BoCF方法对于噪声的鲁棒性归功于本章中采用的DCE方法和形状上下文特征都对噪声具有稳定性。

2.6.7 形状码本大小的影响

在这个实验中，我们来讨形状码本大小（M）对于BoCF方法的性能的影响。实验中学习形状码本的算法固定为k-means。采用的测试数据是整个MPEG-7数据集。使用不同码本大小的情况下，BoCF形状分类准确率如图2-11所示。大致上，形状分

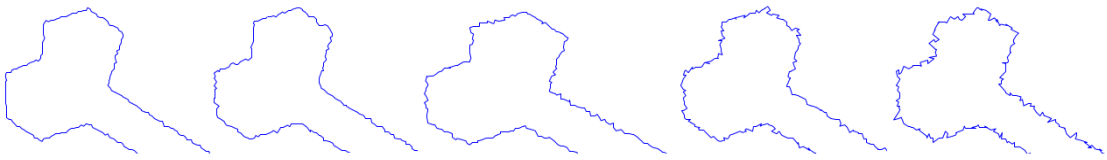


图 2-9 不同程度下的高斯噪声下的形状边缘。从左至右，噪声方差σ依次为0.2，0.4，0.6，0.8和1。

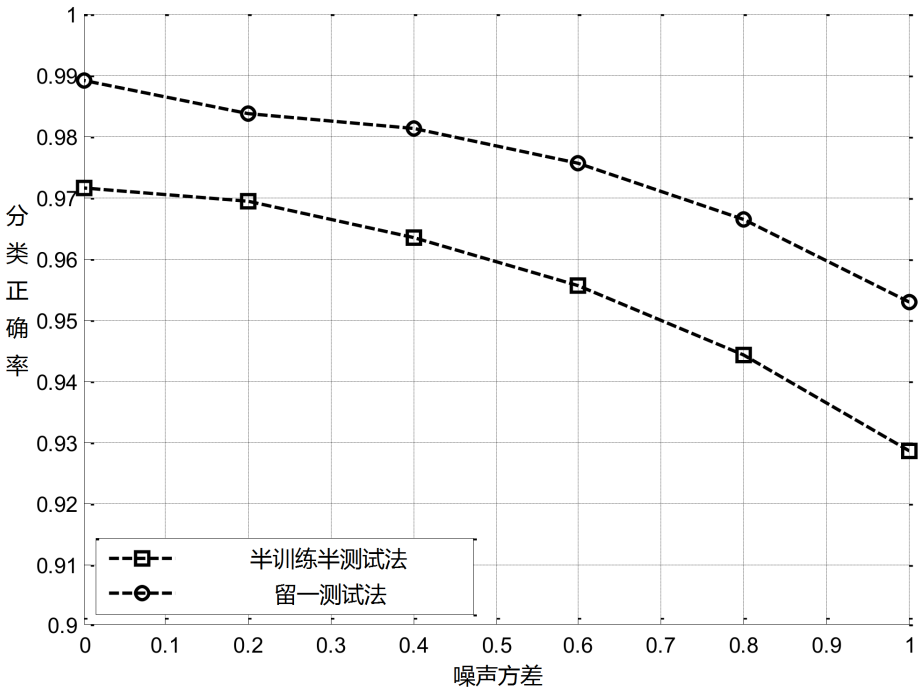


图 2-10 当高斯噪声的方差σ从0变化到1的过程中，在MPEG-7数据集上采用留一测试法和半训练半测试法得到的分类准确率。

类的准确率随着码本大小的增加而提高，但是当码本的大小的增加到1500时候准确率趋于饱和。

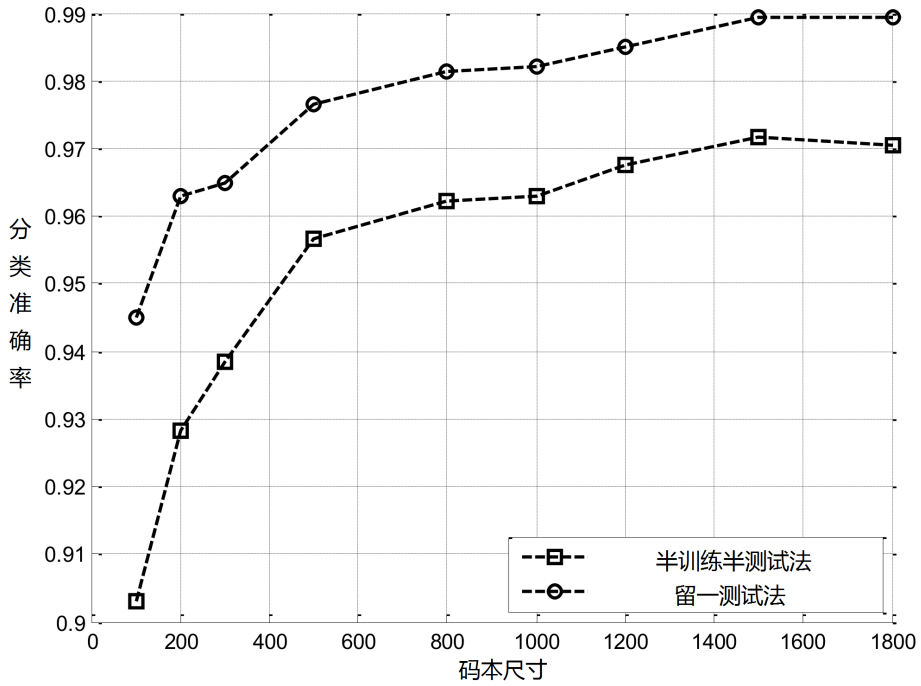


图 2-11 当形状码本从100变化到1800的过程中，在MPEG-7数据集上采用留一测试法和半训练半测试法得到的分类准确率。

2.6.8 形状码本的泛化能力

在这个实验中，我们研究了通过k-means学习的形状码本的泛化能力（generalization ability）。显然形状轮廓片段的空间远小于自然图像的局部特征空间，如SIFT和HOG，那么在这个实验中我们来研究是否有可能学习通用形状码本，这个码本在一个数据集上学习得到。因而，本次实验使用从MPEG-7数据集中学习的码本对动物数据集中的轮廓片段进行编码建立形状并完成形状分类；同样，使用从动物数据集上学习的码本作对MPEG-7数据集中的轮廓片段进行编码建立形状并完成形状分类。规定两个码本的大小都是1500。除了码本之外的所有其它的实验参数都设为一样的。我们可以将这个实验取名为“码本交换”（Codebook Exchange）。表2-6给出了形状分类的结果。结果说明在码本交换之后，在分类准确率上只有大约1%的下降。这些结果表明BoCF方法中的形状码本有着非常好的泛化能力。为什么码本交换之后仍有好的分类效果呢? 原因在于不同的数据集共享着很多相同的轮廓片段。例如，马的腿与狗的腿在轮廓片段上很相似，苹果树的树叶与蝙蝠的翅膀很像。码本交换的成功表明对于所有的基于码本的形状识别系统，可以使用一个通用的形状码本。

表 2-6 码本交换之前和之后的形状分类准确率

	Mpeg-7数据集	Animal数据集
原始方法	97.16±0.79%	83.40±1.30%
码本交换	95.55±0.55%	82.40±1.07%

2.6.9 Caltech 101数据集上的图像分类

之前的众多实验验证本章提出的BoCF方法在形状识别方面优异的精度以及稳定性。那么BoCF方法能否应用于自然图像的分类呢? 这个实验中讨论这个问题。实验中采用被广泛用于测试图像分类算法的Caltech 101数据集[89]。Caltech 101数据集共9144张图像包含101个物体类别以及1个背景类别，这些图像在形状、颜色、纹理上都有很大的变化。每类图像的数量从31到100不等。对于Caltech 101数据集，本次实验中遵循标准的实验设置，对每个类随机选择30张图用来训练，剩下的图像用来测试，对于每一个类，分别计算分类准确率，然后再计算102个类别的平均分类准确率。运行5次，最后计算5次平均的分类准确率作为性能评估的指标。

图 2-12 Caltech 101数据集[89]上的一些图像，(a)和(d)是原始图像，(b)和(e)是gPB算法输出的边缘图像，(c)和(f)是后处理之后得到的形状二值图。

由于Caltech 101数据集中的是彩/灰度图，这与在之前实验中使用的二值形状图不同。因此需要给出对于一个彩/灰度图如何使用BoCF方法建立图像表示的过程：对于一个彩/灰度图，首先使用gPB算法[90]计算它的边缘图（图2-12(b)和(e)展示了一些边缘图），再将所有边缘图上值大于0.1×255的像素设为边缘像素；然后在二值图上利用边缘链接（edge-linking）算法[91]在二值图中寻找一系列的轮廓（如图2-12(c)和(f)所示）。最后用如图2-1所示的步骤(b)-(g)来建立图像表达。与形状分类相同的是，本实验中使用线性SVM进行图像分类。

与SIFT比较 如表2-7所示，使用相同的编码方法（LLC），汇聚方法（SPM），码本大小（1024）及分类器（线性SVM），直接将BoCF里面的轮廓片段特征与文献[14]中的稠密SIFT（dense SIFT）特征进行比较,可直接利用作者发布的源代码运行得到稠密SIFT特征的分类结果。使用LLC和SPM汇聚方法的BoCF准确率为54.5%，比使用LLC和SPM的稠密SIFT特征所达到的71.7%要差。结果较差的原因主要有两个：(1)即使采用目前最好的边缘检测方法来获得彩色图像的边缘图，但物体轮廓上的一些细节，如物体的轮廓(如在图2-12)中车的外形)和物体部分(如图2-12)中人的鼻子)在边缘图中丢失；(2) 一些对于图像分类的十分有效的存在于背景的上下文信息，如车辆图像中的大地和大象图像中的草和树，不能够被轮廓片段特征获取，上述信息采用稠密SIFT可以有效的描述。尽管BoCF比稠密SIFT性能差，但是如表2-7所示，轮廓片段和稠密SIFT特征是互补的。在表2-7中，LLC和RBC[16]两个方法都是基于稠密SIFT的方法。通过使用简单的LP-β方法[92]，将BoCF与稠密SIFT方法结合起来，平均的图像分类准确率可以分别提高3.7%和2.2%。

表 2-7 Caltech 101数据集上的图像分类准确率

方法		平均分类准确率(%)
SVM-KNN[93]		66.2±0.4
SLRR[94]		73.6
LSGC[95]		75.1
稠密SIFT+LLC[14]		71.7±0.8
稠密SIFT+RBC[16]		75.6±0.8
Shape Context[16]		3.0±0.7
BoCF	level 1×1	23.9±0.8
	level 2×1	40.9±0.7
	level 3×3	49.8±0.7
	level 4×4	51.7±1.2
	pyramid	54.5±1.5
	pyramid+稠密SIFT+LLC	75.4±0.8
	pyramid+稠密SIFT+RBC	77.8±1.0

与基于形状的方法对比 首先我们可以设置一个基于形状上下文方法的基准：通过在二值边缘图中设置16个参考点，使用形状上下文特征生成了960维的特征向量，然后使用一个基于形状上下文特征的线性SVM分类器进行图像分类。如表2-7所示，采用形状上下文特征平均图像分类准确率仅有3%。形状上下文和BoCF都是仅采用形状特征的方法。通过将轮廓分成片段然后对其形状上下文特征进行编码，对比直接

采用形状上下文特征，BoCF实现了性能上飞跃，达到了54.5%的分类准确率。这个说明在自然图景中，对于阻挡或者边缘丢失的情况，BoCF比形状上下文描述子有更好的鲁棒性。

金字塔空间的可靠性 表2-7给出从1×1到4×4的不同层面变化后BoCF的准确率从23.9%提高到了51.7%这一结果。通过将四个空间层次结合，BoCF 的准确率是54.5%。这个说明在图像分类中SPM对于BoCF是有显著效果的。

表2-7引用了一些在该数据集上的一些最新结果[94,95]以及一个被称为SVM-KNN[93]的经典方法的结果。综上所述，本实验采用BoCF方法进行自然图像分类进行了全面深入的研究分析，并且说明通过结合BoCF和基于SIFT的方法可以得到比最近的方法[94,95]更好的性能。

2.6.10 时间复杂度分析

轮廓片段包在进行形状识别时十分高效。假定在形状识别时，训练集中的形状数目为m，形状中的局部描述子的数量为n。传统的基于匹配的形状识别方法的时间复杂度一般为$\mathrm{O}\big(mn\log(n)\big)$，每次匹配的典型复杂度为$\mathrm{O}\big(n\log(n)\big)$，$m$个训练样本因此需要进行$m$次匹配。然而，在本章中的方法中，由于可以采用线性SVM分类器，避免耗时的形状匹配。特征编码的时间复杂度为$\mathrm{O}\big(n\big)$，采用线性SVM进行分类的时间复杂度为$\mathrm{O}\big(kc\big)$，其中k为码本大小，c为类别数目，因此，采用轮廓片段包方法进行形状识别的复杂度是$\mathrm{O}\big(n+kc\big)$。

2.7 本章小结

本章提出了一种新颖的形状表示方法：轮廓片段包（BoCF）。本章方法将局部约束的线性编码（LLC）和空间金字塔匹配（SPM）连同词袋（BoW）模型框架引入到形状的表示中。BoCF是基于形状局部特征的，因此它对于形状识别中的形变、局部缺失非常鲁棒。本章针对BoCF方法做出了大量的实验测试，在绝大部分的形状识别测试集上，BoCF达到了目前最高的识别结果。另外，实验中还测试了BoCF 在自然图像进行图像分类的性能，结果显示BoCF明显优于其它形状描述符，并且同基于纹理的图像识别方法互补。在未来的研究中，将继续研究BoCF方法在自然图像中进行物体识别的应用。如将BoCF同滑动窗口方法结合来实现物体检测，或者采用BoCF方法为其它的一些已有的物体检测方法提供形状识别的线索。

3　扇形形状模型

本章介绍一种基于形状的可以用于自然图像中非刚性物体检测的可变形部件模型。物体模型中的每个部件采用形状特征表示，如图3-1中折扇中的转轴，物体的每个部件都可以相对物体模型中的参考点灵活转动，以应付图像中非刚性物体的各种形变，因此该模型被称之为扇形形状模型（FSM，Fan Shape Model）。扇形形状模型采用动态规划算法自动学习得到，在自然图像中进行物体检测时可以快速地自动确定物体的尺寸，并推测出物体轮廓的位置。扇形形状模型可用于形状分类、形状聚类、自然图像中物体检测，均取得了优异的性能。

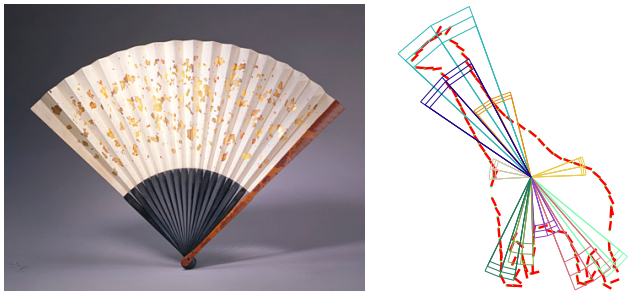


图 3-1　左图显示一把折扇。右图显示利用本章算法学习得到的长颈鹿的扇形形状模型。长颈鹿的每一个部件都被建模为一个灵活的射线。图中用不同的颜色描绘了长颈鹿的100个部件中的10个部件，其中梯形区域显示了每一个部件可以活动的区域。

3.1　研究现状

形状特征对于图像中的光照变化，物体的纹理、颜色变化具有很强的鲁棒性，更能够高效地表达物体大尺度的空间变形，因此采用物体形状特征进行物体识别得到了整个计算机视觉研究领域的高度关注，近年来涌现了大量的基于形状特征的物体识别方法。

在自然图像中提取物体的边缘，即边缘提取（Edge Detection）是基于形状特征进行物体识别的基础。经典的Canny边缘检测算法[96]已经越来越不能够满足如今基于形状的物体识别的需求。Malik等人采用局部光照、纹理、颜色等特征，结合图像分割技术提出Berkeley Edge Detector[97]为基于形状的物体识别方法提供了一个可靠的输入。后续出现了更多的基于有监督学习的边缘检测算法[98–100]，进一步提高了图像中物体边缘检测的精度。尽管如此，目前最好的边缘检测算法输出的边缘图像上均存在大量的物体边缘缺失和背景噪声。

目前，多数的形状的物体检测模型均采用自底向上的形状识别方法。Ferrari等人[39]利用相邻的轮廓对作为特征，来匹配训练集中的物体轮廓和图像中的边缘特征。Shotton等人[40]和Opelt等人[41]几乎同时并分别独立地提出了通过学习一个轮廓片段的码本来进行基于形状的物体识别，他们均采用Chamfer距离[37]作为轮廓间的距离度量，采用Boosting[22]来进行区分训练。沿着文献[40,41]中的思路，Yarlagadda和Ommer[42]提出利用多示例学习来发现有意义的轮廓，采用快速的方向性的Chamfer距离[38]来匹配形状，在ETHZ形状数据集[101]上取得了目前最好的物体检测结果。Srinivasan等人采用一种多对一的匹配方法，并采用具有隐变量的SVM（latent SVM）来保证模型的区分性。在文献[43]中，Ma和Latecki利用部分轮廓匹配[44]来获取物体部件的线索，并采用全局的形状相似性来对识别结果重排序，得到了优异的物体检测性能。另外Toshev等人[102]将检测物体的形状同图像分割相结合，利用将图像分割得到的超像素（superpixel）间的边界作为轮廓，得到了一种新的基于形状的物体检测思路。Yang等人[103]采用一种在由轮廓段构成的图寻找支配子集（dominate set）的方法来实现物体检测。

上述方法均只关注于采用各种方法将训练集中的形状（或轮廓）同测试图像中的边缘匹配起来，缺少一个显性的物体模型，一个既能够抓住物体的结构，也能表示物体外观的模型。缺少一个显性的物体模型的问题在于：当训练样本集中的形状变多，这些自底向上的方法会有更多的轮廓模板需要匹配，导致物体识别速度变慢。因此，部分方法通过区分性学习[40–42]来减少需要匹配的轮廓，但这样会导致物体模型的泛化能力差，例如在一个数据集上训练的模型并不能在其它的数据集上检测同类别的物体。鉴于之前方法的问题，本章中将研究如何从训练样本中的物体轮廓中抽象出结构化的参数化的物体模型。在本人参与的之前的研究工作活动骨架（Active Skeleton）方法[35]中，我们采用物体骨架作为支撑定义了一个结构化的基于形状的物体模型。在活动骨架模型中，物体一般有6个左右的部件表示，容忍部件缺失的能力有限。另外在活动骨架模型中，每个部件是一个轮廓段的集合，并没有将所有的训练样本参数化，识别速度依旧受训练样本数目的限制。本章介绍的扇形形状模型是一个完全参数化的结构化的物体模型。每一类物体模型均采用100个部件，可以容忍不同程度的边缘缺失；每一个采用若干个高斯模型和KNN密度估计表示。在物体检测的实际应用中，扇形形状模型表现出了很强的泛化能力，如图3-2所示。

扇形形状模型是典型的基于部件的物体模型，基于部件的物体表示模型的精髓在于既刻画了物体部件的外观（形状和纹理）又描述不同部件中的空间关系。之前的基于部件的物体模型的研究工作[23,106–108]为扇形形状模型提供了理论支撑。总的来说，扇形形状模型从形状建模的思路为采用基于部件的物体模型提供了一个切实可行的方案。接下来，将给出扇形模型的具体技术细节和实验。

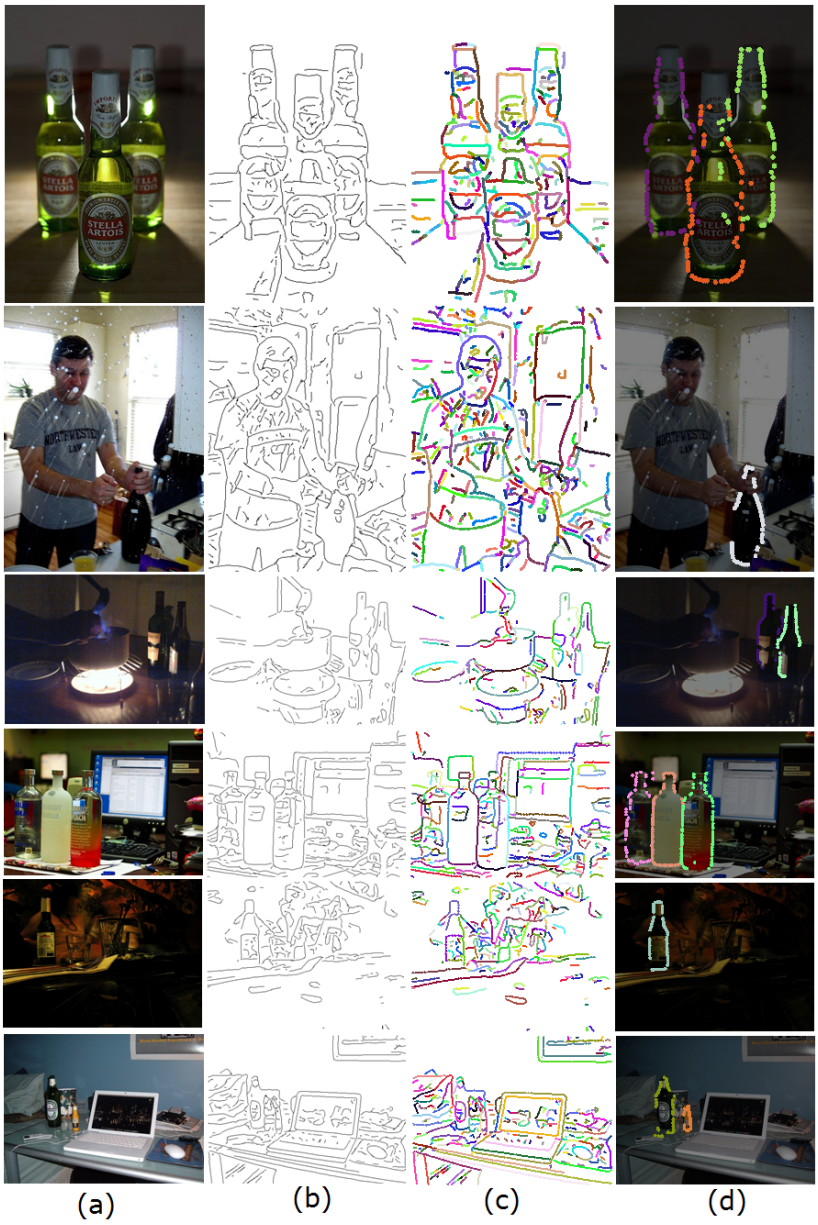


图 3-2 采用ETHZ数据集[101]中瓶子形状训练出的瓶子模型在PASCAL数据集[104]上的瓶子检测结果。（a）显示PASCAL数据集[104]中瓶子图像；（b）显示Berkeley Edge Detector[97]检测得到的边缘图像；（c）现实采用边缘连接[105]处理后得到的结果，可以看出尽管采用了边缘连接，边缘缺失、破损现象还是很严重的；（d）显示采用ETHZ数据集上训练得到的瓶子的扇形形状模型的物体检测的结果。

3.2 基于射线的形状表示和形状匹配

在定义扇形形状模型之前，本章首先提出一种新颖的形状表示方法——基于射线的形状表示（ray-based shape representation），并基于这种形状表示来完成形状的匹配。

3.2.1 基于射线的形状表示

给定一个形状S，采用Canny等边缘检测算法就可以轻松的提取它的轮廓，并在轮廓上均匀取n个有序采样点$\{p_1\ p_2\ \ldots\ p_n\}$。对于这些采样点，可以使用射线$\overrightarrow{op_i}$来描述一个采样点$p_i, 1 \leq i \leq n$，这里的$o$是一个固定的参考点。一般来说，这个参考点$o$可以设置为形状$S$的中心。射线$\overrightarrow{op_i}$的几何模型定义为$R_i(S,o) = (\theta_i, d_i, \alpha_i)$。如图3-3所示，$\theta_i$是射线的倾斜角（它的取值范围为$[0, 2\pi]$）。$d_i$是射线的长度，也就是参考点$o$和$p_i$之间的欧式距离。本文使用形状$S$所有射线的平均长度对其进行归一化。$\alpha_i$ 是轮廓点p_i的边缘方向（它的取值范围是$[0, \pi]$）。我们表示形状S为

$$\{R_i(S,o), i = 1 \ldots n\} \tag{3-1}$$

由于在形状S中我们只设一个参考点o，因此为了表示简洁，我们在下文中使用$R_i(S)$来简洁表示$R_i(S,o)$。

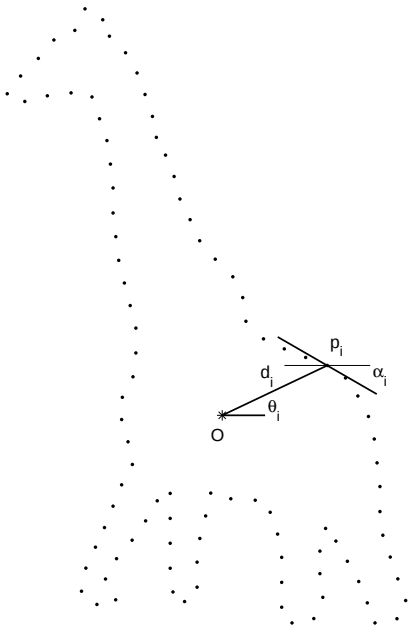


图 3-3 基于射线的形状表示

3.3 基于动态规划算法的形状匹配

基于之前给定的形状表示，我们可以采用$\{R_i(S_1), i = 1 \ldots n\}$和$\{R_j(S_2), j = 1 \ldots m\}$ 分别表示两个不同的形状S_1和S_2，它们的参考点o_1和o_2被省略不写。在此基础上，形状匹配问题定义为寻找一个从$\{i = 1, 2, \ldots, n\}$到$\{j = 0, 1, 2, \ldots, m\}$的映射$\phi$，在此映射函数下，如果$\phi(i) \neq 0$，则$R_i(S_1)$被映射到$R_{\phi(i)}(S_2)$，否则没有$R_i(S_1)$相匹配的射线。形状匹配的目的是寻找一个$\phi$使得匹配代价$C(\phi)$最小化，$C(\phi)$定义如下

$$C(\phi) = \sum_{1 \leq i \leq n} c(R_i(S_1), R_{\phi(i)}(S_2)), \tag{3-2}$$

其中，如果$\phi(i) = 0$，则$c(R_i(S_1), R_{\phi(i)}(S_2)) = \tau$。$\tau$是当射线$R_i(S_1)$没有匹配的惩罚因子。否则，$c(\cdot)$是射线$R_i(S) = (\theta_i, d_i, \alpha_i)$和$R_{i'}(S') = (\theta_{i'}, d_{i'}, \alpha_{i'})$的匹配代价，这个代价定义为

$$\begin{aligned} c(R_i(S), R_{i'}(S')) = \lambda_t * \min(|\theta_i - \theta_{i'}|, 2\pi - |\theta_i - \theta_{i'}|) + \\ \lambda_d * |d_i - d_{i'}| + \lambda_a * \min(|\alpha_i - \alpha_{i'}|, \pi - |\alpha_i - \alpha_{i'}|), \end{aligned} \tag{3-3}$$

其中，λ_t，λ_d和λ_a分别是倾斜角差异，射线长度差异和边缘方向差异的权重系数。

由于形状的轮廓提供了射线的顺序，映射函数ϕ也应当遵循这个顺序。对于这个序列匹配问题，本章中采用经典的动态规划算法来高效的解决。动态规划算法被广泛应用于轮廓匹配问题，并且已经被证明在不同形状间可以取得很稳定的匹配结果。其它的一些序列匹配方法，如ICP[109]、OSB[69]、DTW[110]、编辑距离[62]等，也可以被用于本章的形状匹配中。本章使用文献[111]中的标准动态规划算法，依照公式(3-2)和公式(3-3)中定义的代价函数。在默认情况下，上述匹配过程中两个轮廓的起始点和终点已经对齐。但在实际情况下，这种起点终点对齐的假设是不成立的，所以固定第二个轮廓的起始点（随机选择），尝试将第一个轮廓中的所有采样点作为起始点来进行n次动态规划匹配，最后选择这n次匹配当中匹配代价最小的匹配作为最终的匹配结果。图3-6中给出了两个典型的形状匹配结果，可以看出，本章提出的形状匹配算法对于物体形变非常鲁棒。

3.4 低秩形状与形状方向归一化

回顾本章提出的基于射线的形状表示中，描述一个形状有三个参数，即射线长度d_i，边缘方向α_i和倾斜角θ_i。在这三个参数中，射线长度具有旋转不变性，但是边缘方向和倾斜角这两个特征都对于旋转敏感。这种对于旋转敏感在某些数据集上进行物体检测可能有一定的优势，因为这些物体的方向可能具有一致性。然而，在其它的很多应用场景下，物体可能有着不同的方向。在形状检索和分类上尤其如此。为了在这些应用场景下使用基于射线的形状模型，本文中将对形状进行方向归一化。

在之前的文献中，传统的形状归一化方法皆基于Moment方法。在本章中，我们提出一种新颖的形状方向归一化方法，并将在章节3.7.1用于形状检索。

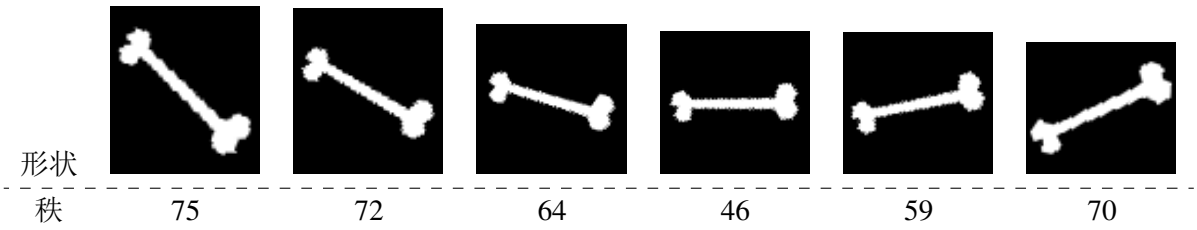


图 3-4 形状的秩随着旋转变化。上: 一个具有不同旋转角度的形状。下: 对应的旋转后的形状的秩。

本文提出的方法从变换不变性低秩纹理（TILT，Transform Invariant Low-rank Textures）[112]中得到了启发。TILT的基本思想是将每一个图像看作一个矩阵，并且在稀疏误差约束下，寻找一个变换来产生一个低秩矩阵。这个低秩矩阵被看作是图像的一个不变结构，对于仿射、投影等变换鲁棒。在本文提出的方法中，我们将每一个形状S看作一个二维矩阵$I \in \Re^{m \times n}$，其中的值是0或者1。形状S的秩定义为$rank(S) = rank(I)$。如图3-4所示，形状S的秩随着形状的旋转而变化，并且旋转到水平角度的形状具有最小的秩46。因此我们需要去发现一个旋转，并且在这个旋转作用下，被旋转的形状具有最小的秩。本文称这个具有最小秩的形状为低秩形状S^*。形状S被旋转$\tau \in [0, 2\pi]$角度记为$S \circ \tau$，并且通过求解下列优化问题来计算S的低秩形状S^*

$$\begin{cases} \tau^* &= \arg\min_\tau rank(S \circ \tau) \\ S^* &= S \circ \tau^* \end{cases} \tag{3-4}$$

这个优化问题是TILT的简化版本，采用增广拉格朗日乘子法[113](ALM，Augmented Lagrange multiplier) 可以用来求解这个全局优化问题。更多关于此优化问题的细节可参考引文[112]。图3-5展示了MPEG-7形状数据集[80]的一些形状方向归一化结果。

图 3-5 形状方向归一化结果：第一行展示了MPEG-7数据集的原始形状；第二行展示了对应的低秩形状。

3.5 扇形形状模型

3.5.1 扇形形状模型的定义

给定一个包含一系列训练样本图像$I_1, I_2, \ldots, I_M$的形状类别$\mathcal{S}$。对于每一个图像，对应的分割后的形状$S_i \subset I_i, i = 1, \ldots, M$也给定。此给定形状类别$\mathcal{S}$的扇形形状模型（FSM，Fan Shape Model）包含一系列的有序扇形部件

$$FSM(\mathcal{S}) = \{F_i, i = 1 \ldots N\}. \tag{3-5}$$

其中F_i表示$\mathcal{S}$的第i个扇形部件。不同于章节3.2.1中的射线，每一个扇形部件是一组参数的分布，而不是单独的一个值。扇形部件F_i定于如下：

$$F_i = (\mu_\theta^i, k_\theta^i, \mu_d^i, \sigma_d^i, \Lambda^i). \tag{3-6}$$

这个分布包含两类参数。一类是用以建模物体部件的空间信息，比如μ_θ^i和k_θ^i描述了射线i角度的分布，μ_d^i和σ_d^i描述了射线i的长度分布。另一类参数用以表示物体部件的外观特征，记为Λ^i，其中包含形状信息和纹理信息。关于这两类参数的具体细节以及如何估计这些分布的参数将在下面的章节具体介绍。

3.5.2 扇形形状模型的参数估计

为了估计扇形形状模型的参数，本文将从形状匹配着手。首先，我们需要选择第一个训练形状S_1和它的参考点[⑤]。对于剩下所有训练形状$\{S_2, \ldots, S_M\}$，参考点将通过形状匹配自动决定。对于每一个训练形状，本文在其中均匀设置很多参考点。然后对于每一个参考点通过章节3.3中介绍的动态规划算法来进行与第一个形状S_1的匹配。除了动态规划，由于模型学习的过程中需要学习射线的倾斜角的方差以建模扇形部件的可活动区域，因此设置公式(3-3)中倾斜角的权重系数$\lambda_t = 0$。当所有的参考点都被尝试匹配后，我们选择具有最小匹配代价的参考点。图3-6展示了一个匹配范例。如图3-6所示，本文选择的特征通过动态规划算法，即使在两个长颈鹿形状之间存在大量的变化，也可以提供很好的形状匹配，这为正确估计模型参数提供了条件。

经过形状匹配，我们对于每个射线独立的估计参数。对于S_1中第i条射线($1 \leq i \leq N$)，在$S_2, \ldots, S_M$中一共有$M-1$条与之匹配的射线。因此，在训练集中一共有M条射线属于物体的第i个扇形部件，这些射线表示为$\{R_i(S_1), \ldots, R_{\phi(i)}(S_j), \ldots, R_{\phi(i)}(S_M)\}$。注意到前文所述，对于$1 \leq j \leq M$，$R_{\phi(i)}(S_j) = (\theta_{\phi(i)}, d_{\phi(i)}, \alpha_{\phi(i)})$。

⑤ 实际中，训练形状可随机排序，第一个形状中的参考点设置为它的中心。不同的形状顺序对最终得到的物体模型影响不大

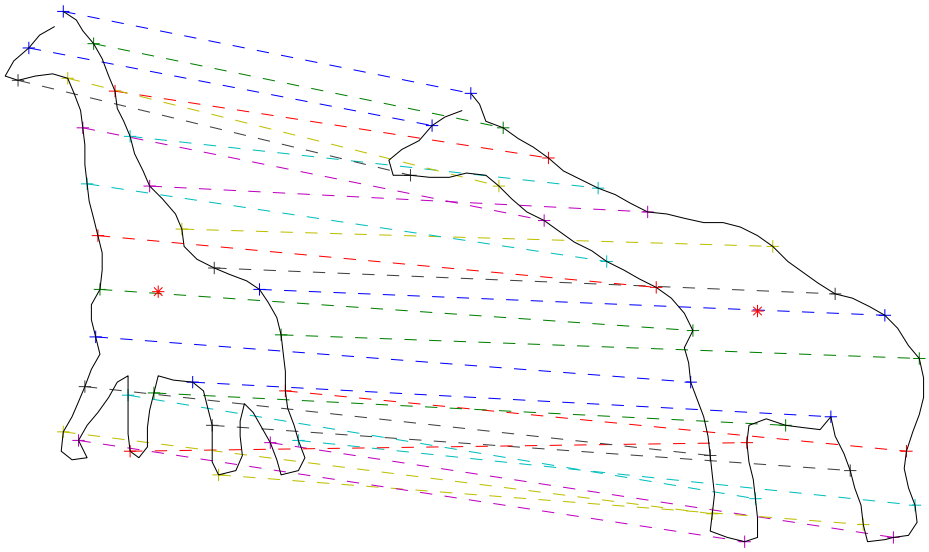


图 3-6 形状匹配结果。红星标记参考点，彩色虚线连接了两个形状的匹配点。

本章认为射线的长度$d_{\phi(i)}$服从高斯分布$\mathcal{N}(d_{\phi(i)}, \mu_d^i, \sigma_d^i)$，倾斜角$\theta_{\phi(i)}$的分布$\mathcal{M}$服从角度域的高斯分布，即冯·米塞斯（von Mises）分布[114]，

$$\mathcal{M}(\theta_{\phi(i)}, \mu_\theta^i, k_\theta^i) \propto e^{k_\theta^i cos(\theta_{\phi(i)} - \mu_\theta^i)}$$

其中，μ_θ^i表示所有$\theta_{\phi(i)}$的均值，k_θ^i是$\theta_{\phi(i)}$集中度的度量。μ_d^i和σ_d^i通过极大似然来估计。类似的，对于冯·米塞斯分布的极大似然参数$(\mu_\theta^i, k_\theta^i)$也有已知的方法。

本章使用两种局部特征作为Λ_i：一种是边缘方向，另一种是提取自局部小区域的SIFT特征[1]。他们分别描述了物体部件的形状和纹理。因此，我们表示Λ_i为

$$\Lambda_i = (\mu_\alpha^i, k_\alpha^i, T^i)$$

边缘方向α取值范围是$[0, \pi]$，2α的分布是冯·米塞斯分布$\mathcal{M}(\alpha, \mu_\alpha^i, k_\alpha^i)$。

本章使用SIFT特征去估计第i个扇形部件的外观分布。T^i是训练集中所有与第i个扇形部件匹配的图像区域的SIFT特征的集合。因此，T^i表示了一个SIFT特征离散的经验分布。固定尺寸的局部小区域被用来计算SIFT特征。对于T^i，本文通过KNN密度估计，来计算一个从测试图像的局部SIFT小区域属于第i个部分的概率。

图3-7展示了本章方法在ETHZ数据集[101]的五个类别上训练得出的模型。从左到右分别是苹果、瓶子、长颈鹿、杯子和天鹅的扇形形状模型。图片展示了一些物体部件的空间分布和边缘方向特征，细节请参考图注。

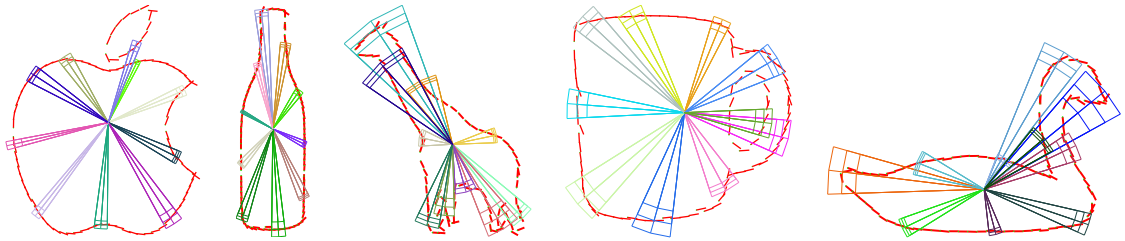


图 3-7 ETHZ数据集训练得出的模型。图中只展示一部分扇形部件，每一条中线的末端表示一个部件由μ_d^i和θ_d^i给定的平均位置，末端的红色线段显示平均边缘方向，也就是μ_α^i。不同颜色的扇区表示不同的物体部件，对于每一个扇区，它的张开角度显示了倾斜角的方差k_θ^i，前端长度变化显示了射线长度的方差σ_d^i。

3.6 基于扇形形状模型的物体检测

在完成对于每一个物体类别自动学习得到物体类别的扇形形状模型之后，我们可以采用这个扇形形状模型在自然图像中来进行物体检测。使用扇形形状模型进行物体检测包含以下两个关键：(1)在给定图像位置处快速估计物体尺寸。(2)在估计的尺寸条件下得出检测假设得分。这两步将在图片的每一个位置进行，本文将在下文详细讲解这两步。经过非极大值抑制（NMS，Non Maxima Suppression），可以得到最终的检测结果。

3.6.1 快速物体尺寸估计和检测假设

在已经学习了一个给定形状类别$\mathcal{S}$的扇形形状模型以后，物体检测的思路是将测试图片和扇形形状模型进行匹配。首先，基于扇形形状模型，本章提出了一个新颖的快速物体尺寸估计的方法，这个方法显著不同于之前的简单的通过变换模型或者测试图片的尺寸，然后在这些尺寸间进行寻找最佳尺寸的方法。给定一个测试图像I，首先计算I的边缘图E，并对于每一个边缘像素，在给定尺寸上计算边缘方向和SIFT特征向量。然后，对于图像I中给定的参考点位置l，考虑物体中所有到l的距离小于最大期望半径r的边缘像素点。这些边缘像素点和它们的特征表示为$Es = \{d_j, \theta_j, \alpha_j, t_j; 1 \le j \le ne\}$，其中$d_j$和$\theta_j$是第$j$个边缘像素$e_j \in Es$相对于$l$的射线长度和倾斜角，$\alpha_j$是边缘方向，$t_j$是在第$j$个边缘像素的SIFT特征向量。$ne$是$Es$中边缘像素的个数。边缘像素$e_j$在尺寸$s$下，属于扇形$F_i$的概率为

$$\begin{aligned} p(e_j|F_i, s) = \mathcal{N}(d_j/s, \mu_d^i, {\sigma_d^i}^2) \cdot \mathcal{M}(\theta_j, \mu_\theta^i, k_\theta^i) \\ \cdot \mathcal{M}(2\alpha_j, \mu_\alpha^i, k_\alpha^i) \cdot NN(t_j, T^i) \end{aligned} \tag{3-7}$$

其中，$NN(t_j, T^i)$是KNN密度估计输出的概率，KNN密度估计使用T^i作为正训练集样本，t_j作为测试样本。为了估计在l处的物体尺寸，本章建立一个离散的尺寸空间S_s和一个与S_s同大小的投票空间V_s。尺寸空间包含ns个可能的尺寸$S_s =$

$[s_1, ..., s_{ns}]$。V_s定义为：

$$V_s[m] = \prod_{i=1}^{N} (\max_j (p(e_j|F_i, s_m))), \tag{3-8}$$

其中$m = 1, \ldots, ns$，N是扇形形状模型的扇形部件个数。在l处最佳的尺寸s^* 为

$$m^* = \arg\max_m (V_s[m]) \text{ and } s^* = s_{m^*} \tag{3-9}$$

因此，对于每一个位置l，都可以确定一个最佳的尺寸s^*，并且这个最佳的尺寸会被用在章节3.6.2描述的检测过程中。

公式(3-8)中提到的V_s可以快速地估算出来：对于边缘像素$e_j = \{d_j, \theta_j, \alpha_j, t_j\}$，首先，没有必要对于每一个$F_i$去计算$p(e_j|F_i, s_m)$。根据正态分布的概率分布，给定$F_i$和$e_j$，如果$|\theta_j - \mu_\theta^i|$大于三倍的$\theta_i$的标准差，$p(e_j|F_i, s_m)$将近似为0。因此，对于这样的$F_i$，$p(e_j|F_i, s_m)$将直接设置为0。其次，给定$e_j$，对于$F_i$，也没有必要在每一个尺寸$s_m$去计算$p(e_j|F_i, s_m)$。只有那些值落在$\frac{d_j - 3\sigma_d^i}{\mu_d^i}$和$\frac{d_j + 3\sigma_d^i}{\mu_d^i}$, $p(e_j|F_i, s_m)$之间的s_m才需要去计算，否则可以直接设置概率为0。按照以上计算方法，可以发现在尺寸变化步长固定的前提下，计算V_s的代价并不会随着尺寸空间的大小变大而变大。

对于每一个位置l和它的最佳尺寸s^*，接下来将介绍如何获得一个检测假设（Detection Hypothesis）。所谓检测假设是指在图像特定区域最有可能是区域、轮廓或者像素等。在本章中检测假设由一组边缘像素$e^*(l) = (e_1^*, ..., e_N^*)$组成，并且满足

$$e_i^* = \arg\max_{e_i} (p(e_i|F_i, s^*)) \tag{3-10}$$

这些边缘像素是在估计到的尺寸下与物体模型中扇形部件最相匹配的轮廓点。最终的检测假设的评估将在下面一个小节介绍。

3.6.2 检测假设评估

在确定物体检测假设的过程中，按照通过公式(3-10)，对于每一扇形部件F_i可以选取边缘像素e_i^*。因为每个扇形部件是独立的，它们之间没有相互制约关系，所以这个步骤计算起来非常高效。实际上，扇形形状模型不仅对于每一个扇形部件计算其空间分布和纹理信息，并且编码所有扇形部件的顺序和它们的相对位置关系。本章提出**支撑轮廓**和相邻扇形部件间的距离一致性两个方法在最终的检测假设评估上使用了扇形部件邻近关系和空间分布信息来进一步提高物体检测的精度。

支撑轮廓： 如果两个相邻扇形部件选取的边缘像素在同一个轮廓片段上，那么这更有可能是一个正确的检测，否则这个假设可能只是因为杂乱背景造成的偶然匹配。

根据这个假设，支撑轮廓定义一个分数：

$$\eta_1 = \frac{\Sigma_{i=1}^{N-1} slen(e_i^*, e_{i+1}^*)}{\Sigma_{i=1}^{N-1} len(e_i^*, e_{i+1}^*)}. \tag{3-11}$$

sp表示e_i^*和e_{i+1}^*的连线$\overline{e_i^* e_{i+1}^*}$上的样本点，$len(e_i^*, e_{i+1}^*)$表示$sp$中点的个数，$slen(e_i^*, e_{i+1}^*)$表示$sp$上具有下列约束的点的个数，即在$sp$周围，至少有一个边缘像素到它的距离小于一个阈值。这个阈值设置为2个像素。注意到断裂的边缘也可以“支撑”sp，所以我们的检测方法对边缘断裂具有鲁棒性。

相邻扇形部件间的距离一致性： 由于在训练过称中，物体轮廓是等距离取样本的，相邻部分的距离应该是一致的。相邻扇形部件的距离序列表示为$d_{neib} = [|\overrightarrow{e_1^* e_2^*}|, ..., |\overrightarrow{e_{N-1}^* e_N^*}|, |\overrightarrow{e_N^* e_1^*}|]$。则距离一致性得分定义为

$$\eta_2 = \exp(-std(d_{neib}/s^*))$$

在l处最终的检测假设得分定义为

$$score(l) = \eta_1 \, \eta_2 \prod_{i=1}^{N} (p(e_i^* | F_i, s^*))$$

3.7 实验

在这个章节，本文将展示扇形形状模型在标准形状数据集MPEG-7[80]上的形状检索表现，在ETHZ数据集[101]上的形状聚类和物体检测性能表现。

3.7.1 形状检索

MPEG-7数据集[80]在过去十年来被广泛用于测试形状描述子的性能，一直被认为是一个标准的形状测试数据集。这个数据集包含1400个二值形状，共分为70类，每类包含20个物体。这是一个十分具有挑战的数据集，其中很多属于同一类形状有着非常复杂的形变。图3-8展示了该数据集的一些范例形状。MPEG-7数据集的检索准确率通常叫做bull's eyes得分，这个得分计算前40个最相似的物体中与待查询物体属于同一类的物体个数除以20（因为每一类有且仅有20个物体）。数据集的所有形状都将用于查询，整个数据集的检索结果是取所有形状的检索结果的平均值。

在这个实验中，所有形状的方向都经过章节3.4描述的方法进行旋转归一化。每一个形状上均匀的采样100个点用于建立基于射线的形状模型；每一个形状的参考点设置为该形状的中心；$\lambda_t = 0.5, \lambda_d = 3, \lambda_\alpha = 1.5$。表3-1比较了本章中提出的基于射线的形状描述子与其它描述子的bull's eyes得分。基于射线的形状描述子

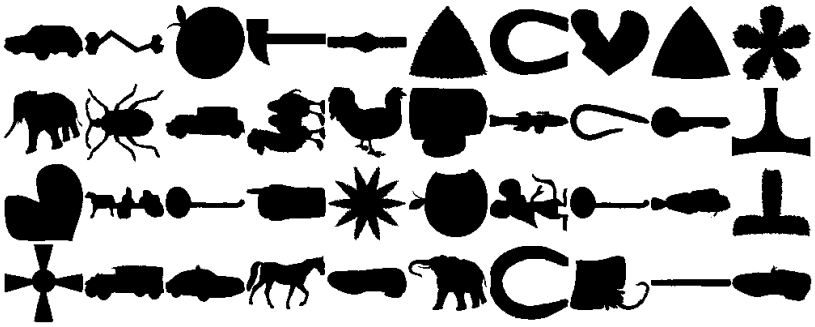


图 3-8 MPEG-7数据集的范例形状

的得分是85.80%，与其它流行的描述子，比如形状上下文(SC)[7]，内距离形状上下文(IDSC)[34]，和形状树[36]的检索结果不相上下。

在所有比较的方法中，本章中的形状描述子是最简单的一个。计算一个基于射线的形状描述子只需要对每一个轮廓采样点计算它的倾斜角，射线长度，边缘方向。单个描述子最好的表现是89.66%[88]。

表 3-1 MPEG-7数据集上Bull’s eyes得分

形状描述子	Bull’s eyes得分(%)
Height Function[88]	89.66
Contour flexibility[59]	89.31
Shape tree[36]	87.70
SC + DP[115]	86.80
IDSC + DP[34]	85.40
SC + TPS[7]	76.51
基于射线的形状表示	85.80

3.7.2 形状聚类

ETHZ形状数据集[101]包含5个不同的类别（苹果，瓶子，长颈鹿，杯子，天鹅），一共有255张图片。每一个类别包含32至86张图片。不同类之间有巨大的尺寸、光照差异。很多物体周围都存在大量的杂乱背景，并且有内轮廓。根据在此数据集[116,117]进行物体检测的一般设定，对于每一类的前一半图片作为训练集，剩下的作为测试图片。数据集内提供的人工标注的物体轮廓用以训练。

在这个实验中，我们假设ETHZ形状数据集所有的训练形状的类别信息未知，希望通过聚类来将这些形状划分到对应的类别中去。为了解决这个问题，具体的解决

方法是根据基于射线的形状描述子，计算任何两个形状之间的距离以获得一个距离矩阵，在距离矩阵上使用关系传播聚类算法（Affinity Propagation）[118]，把这些形状分为不同的类别。图3-9展示了上述方法的聚类结果。如图所示，只有一个存在着严重遮挡的长颈鹿被错误分类到瓶子类别里（第二行最右面那个）。所有其它的聚类结果都是正确的。这结果显示，给定一组来自不同类别的形状，本章的方法可以同时将这些形自动的状分为不同类，并且对每一个类独立学习物体模型。然后，这个学习到的物体模型可以接下来被用在新的图像上进行物体检测。

图 3-9 ETHZ数据集所有训练形状的聚类结果。每一行显示一个类的形状。

3.7.3 物体检测

在这个实验中，如章节3.5所述，我们对ETHZ数据集中5个不同的类别分别学习扇形形状模型，然后按章节3.6描述的方法进行物体检测。图3-7展示了学习的扇形形状模型。每个扇形部件的纹理特征利用数据集中给定的灰度图上提取的SIFT特征。

本文使用Berkeley Edge Detector[97]在测试图片上去提取概率边缘图。为了将概率边缘图转换为二值边缘图，概率边缘图所有像素值大于0.02的像素均被视为边缘像素。这意味着实验中并不会针对不同的图像去调整这个阈值来得到更好的边缘图像。在检测过程中，由于本章提出的物体检测的算法时间复杂度并不随着尺度空间尺寸的增大而增大，因此一共有15个尺寸被采用。最后采用非极大值抑制方法来去除冗余检测假设。

为了评估检测结果，实验中采用PASCAL标准，也就是当检测的包围盒（Bounding Box）和人工标注的物体包围盒的交集与两个包围盒的并集的比值大于50%，这个检测就认为是正确的。

实验中将基于扇形形状模型的物体检测方法与一些流行的基于轮廓的物体检测方法[116,117]以及一些基于纹理的区分式部件模型[120]进行比较。所有这些方法和本章的方法使用相同的测试集和训练集。图3-10展示了精度/召回率曲线（PR，Precision/Recall），平均精度（AP，Average Precision）采用使用工具包[104]来计算，表3-2展示了其它方法[42,117,119–121]的AP值。本章的方法的平均AP是0.869。最高的平均AP值是通过多示例学习来学习有意义的轮廓和发现这些轮廓的相互位置得到

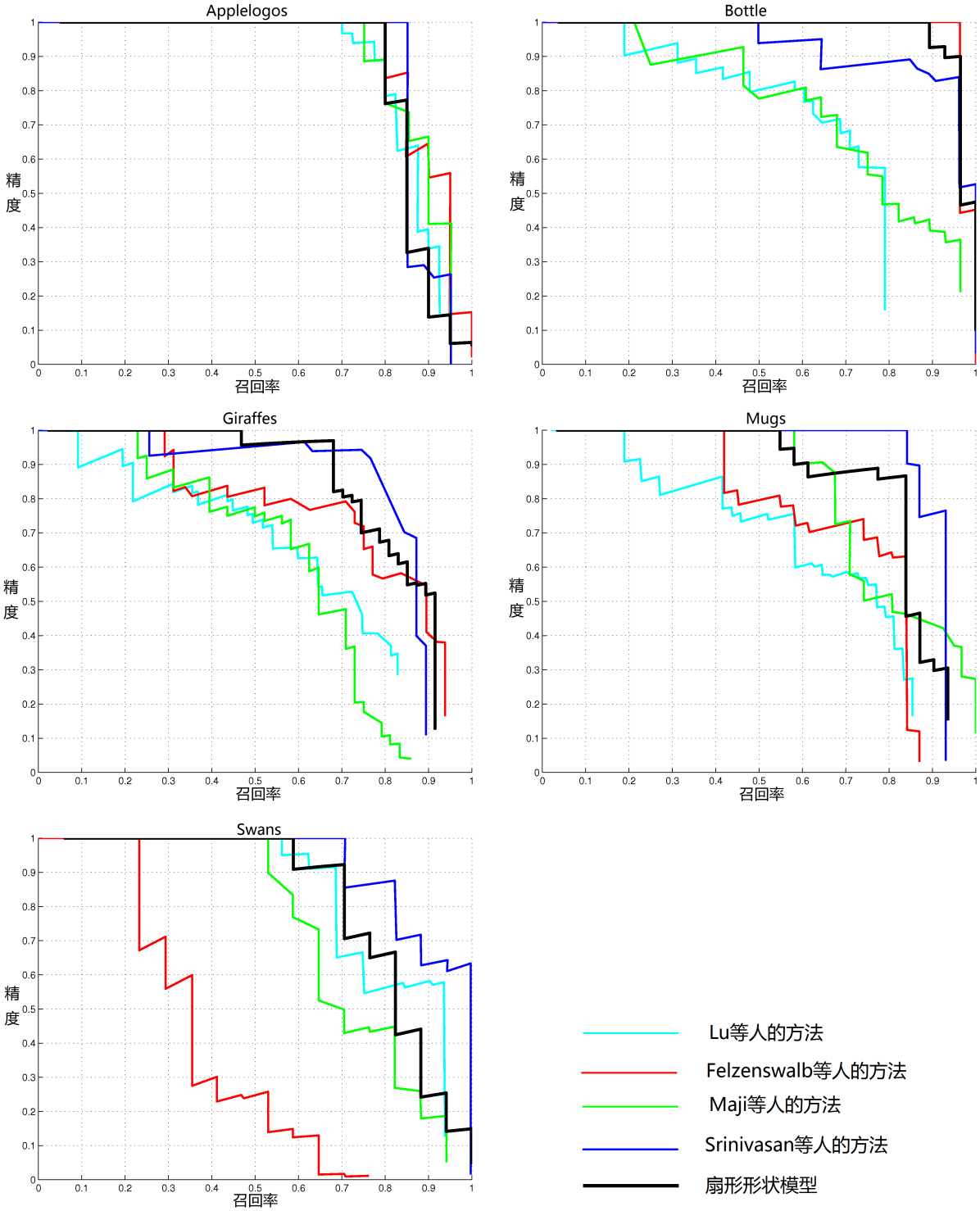


图 3-10 ETHZ形状数据集上，扇形形状模型和Lu等人[121]，Felzenswalb等人[120]，Maji等人[119]和Srinivasan等人[117]的方法的精度/召回率曲线比较。

表 3-2 ETHZ形状数据集的平均精度(AP)比较

	Applelogos	Bottles	Giraffes	Mugs	Swans	Mean
扇形形状模型	0.866	**0.975**	**0.832**	0.843	0.828	0.869
Yarlagadda等人的方法[42]	-	-	-	-	-	**0.882**
Srinivasan等人的方法[117]	0.845	0.916	0.787	**0.888**	**0.922**	0.872
Maji等人的方法[119]	0.869	0.724	0.742	0.806	0.716	0.771
Felzenszwalb等人的方法[120]	**0.891**	0.950	0.608	0.721	0.391	0.712
Lu等人的方法[121]	0.844	0.641	0.617	0.643	0.798	0.709

的0.882[42]⑥。在这些方法中，扇形形状模型在长颈鹿大类里取得最高的AP值，这个类别对其它方法来说，是该数据集最困难的一个类别。值得注意的是扇形形状模型超越了流行的可变形部件模型（DPM，Deformable Part Model）方法[120]。扇形形状模型和DPM都是基于部件的方法，而这些部件都采用了纹理特征来描述。结果显示扇形形状模型比其它基于部件的物体模型[120]更加灵活。

表 3-3 ETHZ形状类别在0.3/0.4 FPPI检测率的比较。

	Applelogos	Bottles	Giraffes	Mugs	Swans	Mean
扇形形状模型	0.90/0.90	**1/1**	**0.92**/0.92	0.94/0.94	0.94/0.94	0.94/0.94
Yarlagadda等人的方法[42]	**0.95/0.95**	**1/1**	0.91/0.91	**0.97/0.97**	**1/1**	**0.97/0.97**
Srinivasan等人的方法[117]	**0.95/0.95**	**1/1**	0.87/0.89	0.94/0.94	**1/1**	0.95/0.96
Kontschieder等人的方法[122]	0.94/1	**1/1**	0.91/**0.93**	0.81/0.87	**1/1**	0.93/0.96
Maji等人的方法[119]	**0.95/0.95**	0.93/0.96	0.89/0.89	0.94/**0.97**	0.88/0.88	0.92/0.93
Felzenszwalb等人的方法[120]	**0.95/0.95**	**1/1**	0.73/0.73	0.84/0.84	0.59/0.65	0.82/0.83
Lu等人的方法[121]	0.9/0.9	0.79/0.79	0.73/0.77	0.81/0.83	0.94/0.94	0.84/0.85
Riemenschneider等人的方法[122]	0.93/0.93	0.97/0.97	0.79/0.82	0.85/0.86	0.93/0.93	0.89/0.91
Ferrari等人的方法[116]	0.78/0.83	0.79/0.82	0.399/0.445	0.75/0.8	0.63/0.71	0.67/0.72
Zhu等人的方法[123]	0.80/0.80	0.93/0.93	0.68/0.68	0.65/0.74	0.82/0.82	0.78/0.79

图3-11展示了每图平均虚警率（FPPI，False Positives Per Image）vs.检测率（DR，Detection Rate）。表3-3比较了在0.3/0.4 FPPI扇形形状模型的检测率和其它方法[116,117,119–123]。扇形形状模型和Srinivasan等人的方法[117]结果不相上下。可以发

⑥ 该方法发表于扇形形状模型之后，在扇形形状模型发表的时候，扇形形状模型在ETHZ数据集上取得了当时最好的物体检测性能。

现扇形形状模型是唯一一个在0.3 FPPI和0.4 FPPI上的检测率没有差异的。扇形形状模型的曲线在最开始急剧上升，然后在0.3/0.4 FPPI前达到检测率的最高点。

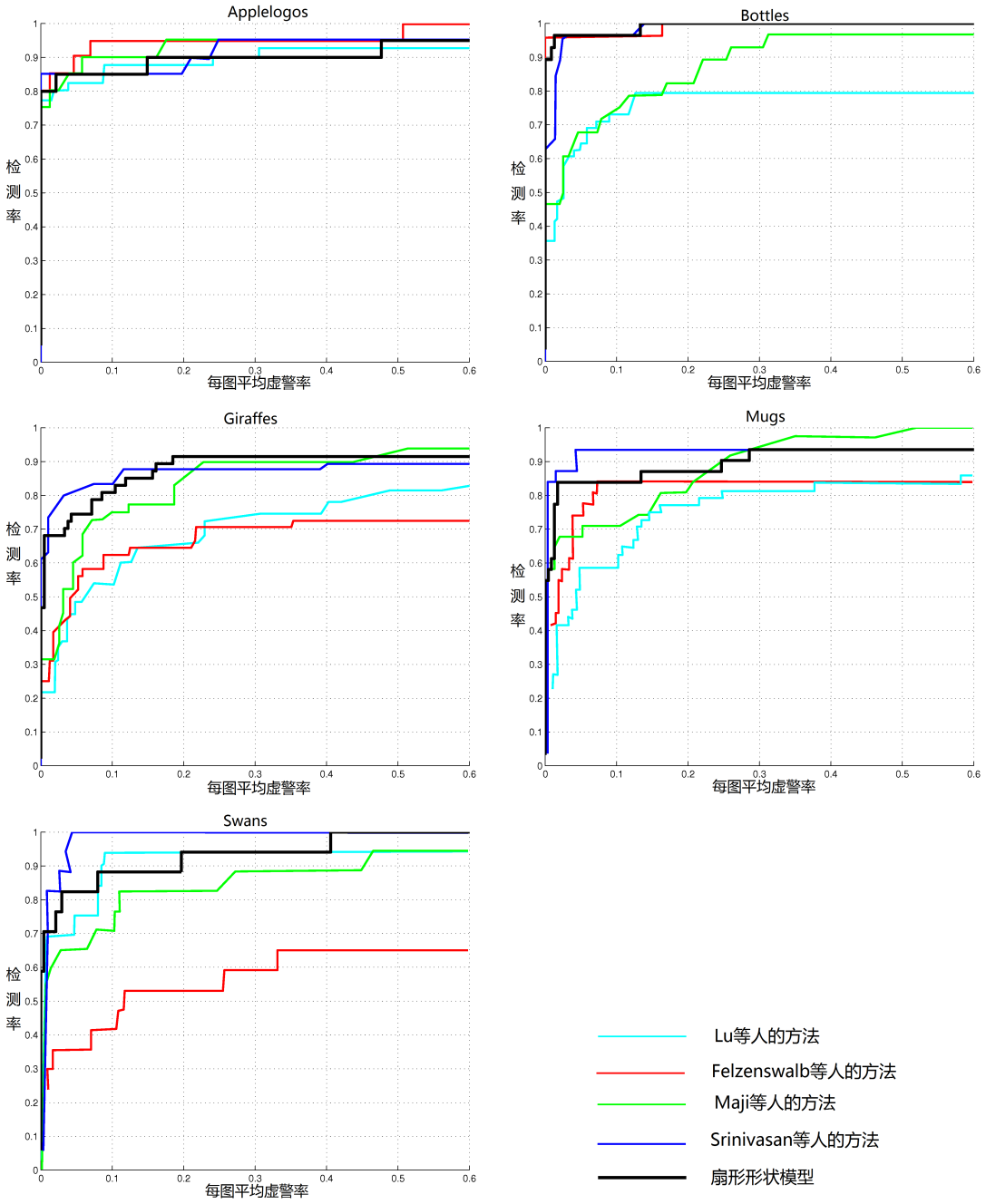


图 3-11 ETHZ数据集上，扇形形状模型与Lu等人[121]，Felzenswalb等人[120]，Maji等人[119]和Srinivasan等人[117]的检测率/每图平均虚警率曲线比较。

图3-12展示了扇形形状模型的部分检测结果，包括正确检测和错误检测。内轮廓和外轮廓（例如，杯子手柄）可以被检测到。可以看出，扇形形状模型可以在嘈

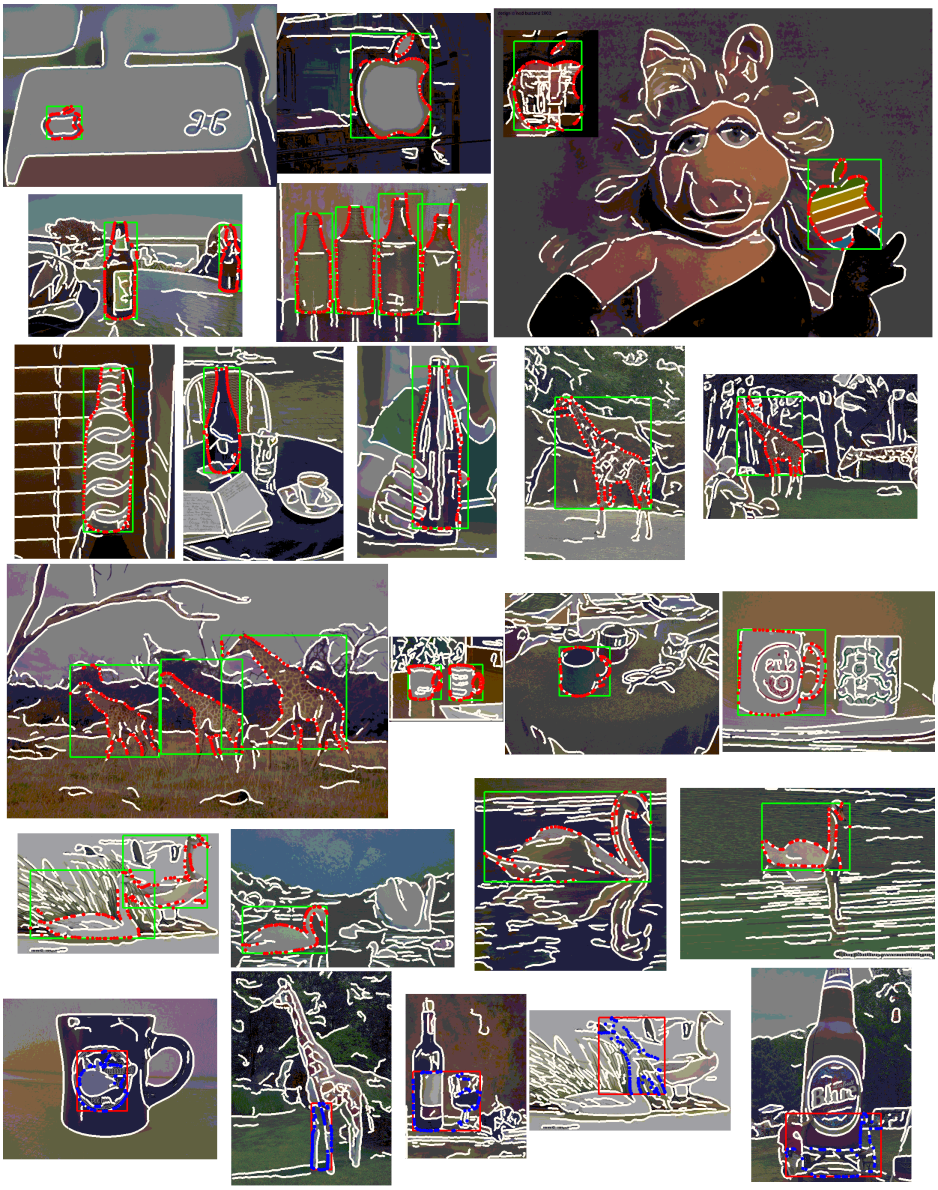


图 3-12 ETHZ形状数据集的部分检测结果。每一张图片展示了一个或多个物体检测的边缘像素和物体包围盒。最下面一行显示了苹果，瓶子，长颈鹿，杯子和天鹅的错误检测结果（从左到右）。边缘像素为白色，在绿色边界框内和红色点上为正确检测，红色边界框内和蓝色点上为错误检测。

杂的物体边缘图像中找出那些尺度极小、边缘破损、形变大的物体。

3.8 本章小结

本章的研究通过从同类物体的轮廓之间的匹配关系抽象出一个基于部件的物体模型——扇形形状模型。该模型的基础是一个简洁的基于射线的形状表示。扇形形状模型利用高斯分布、KNN密度估计等一系列参数化的表示方法刻画了物体部件的灵活的空间分布和丰富的外观特征，在物体检测应用中具有能克服物体形变、边缘缺失等优点。今后的研究重点将是在物体检测中精确推断扇形形状模型以及如何提取更好的提取图像中的物体边缘。

4 基于低秩优化的子空间发现

本章研究如何从存在大量离群点和噪声点的数据中鲁棒地发掘子空间的问题。一个典型的例子是：给定一系列图像，其中每一幅图像中都包含一张人脸，但人脸所处的位置和大小均是未知的。本章的目标是自动地去发掘出人脸在图像中所在的位置，并学习得到人脸的外观模型。针对这个十分困难的问题，本章从一个全新的方向进行探索，基于简洁的生成型模型和低秩优化方法，提出一套可以同时定位物体并学习物体的外观模型的数学框架。针对其中的数学优化问题，本章提出了一种基于交替方向乘子法（ADMM，Alternating Direction Method of Multipliers）的解决方法，并且通过大量的仿真和实验来验证本章方法的有效性。另外，本章提出的方法也可以应用到相关的高维数据的组合优化问题上。

4.1 研究现状

近年来，子空间学习方法已经被应用到诸多的高维样本的分析的问题[124–126]当中。在假设样本在对齐好了且存在于一个低维的子空间的情况下，这些方法可以处理样本中存在的稀疏的大的错误，并且学习得到这个地址空间。其它的方法[127–130]可以将样本聚类到不同的子空间中。但是，对于样本中具有大量的离群点的情况下，这些方法均不能弱监督地学习得到数据中的子空间。图4-1中展示了本章问题的基本设定，同时也展现了本章所提出的解决方案。这里，给定的是一个图像集合，且在这些图像中包含同一个物体（或者模式）。本章的目的是自动地去寻找到物体并学习得到它的子空间模型。

从一个数学抽象的角度来看，给定的是样本中既包含内点和大量的离群点，这些内点存在于一个维数相对较低的子空间中，另外，内点中存在稀疏的错误。借用多示例学习（MIL，Multiple Instance Learning）的概念，给定两个约束，（1）样本被称作是示例（instance），分布在不同的包（bag）中，（2）每个包中存在至少一个内点。这两个约束往往是同时成立的，如图4-1(a)和(b)所示，我们可以将每一幅图像看作是一个包，包中的每一个图像块看作是一个示例，若图像块对应于物体，则为内点，否则为离群点。本章的目的是在这些样本中寻找一个内点所在的子空间。显然，这是一个高度复杂的组合优化问题。这里我们可以借用多示例学习的概念，但是假设在训练过程中不存在负的包。这个设定类似于Zhu等人文章中的设定[131]。原始的问题就可以转化为一个弱监督情况下的子空间发现问题。然后，将这个子空间发现问题进一步转化到一个凸优化（convex optimization）数学框架下，并且采用交替方向乘子法[132,133]来解决这个问题。在本章所提出的数学框架下，每一个示例对应一个表示该示例为内点还是离群点的指示（如图4-1(b)所示）。这些示例的指示被视作

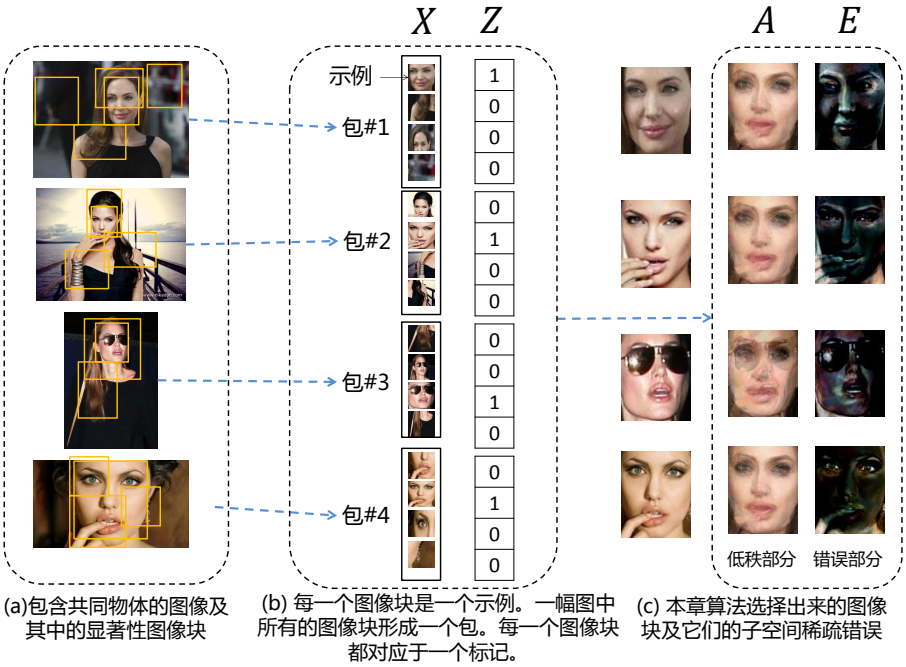


图 4-1 物体发现过程的系统框架图。给定一个图像集合，首先检测其中具有显著性的区域（用矩形标注）。将一幅图像看作为一个包（bag），该图像中的每一个显著性图像块都被看作是包中的一个示例（instance）。假定所要发掘的共同物体作为一个示例存在于包中。然后本章的算法检测到共同物体并学习物体对应的子空间模型，另外除了物体的低秩部分之外，还可以得到物体上的残余部分（即，错误部分）。图上方的符号X, Z, A和E对应于章节4.2中的数学表示。

为隐变量。本章的目标方程通过同时最小化子空间的秩和样本误差的ℓ_1范数来选择示例。这样，通过解决这个凸优化问题便可以实现子空间发现的目的并辨别出哪些示例存在于这个子空间。如图4-1(c)所示，我们可以发现人脸所在的子空间，并辨别出哪些图像块是人脸，给出人脸上存在的误差。在本章的实验中，将展示不仅仅是人脸上的实验结果，还将展示本章的方法可以去发现各种不同类别的物体。

从有噪声的数据中鲁棒地学习出数据模型是机器学习中的核心问题，多示例学习（MIL）[134]就是其中的一种情形，在MIL中，输入数据是以包的形式给定的，正包中含有至少一个正的示例，负包中全部都是负的示例。这是一种弱监督学习的设置。MIL中包含两个核心的子问题：（1）推测示例的标记信息，和（2）根据示例标记学习出正确的数据模型。EM[135]算法经常被用来处理这类标记缺失的问题，类似的还有隐变量SVM（latent SVM）方法[136]。其它的一些方法可以在EM、隐变量SVM的基础上进行改进，但是通过对比可以发现，这些方法往往倾向于贪婪地迭代地寻找到一些次优解（suboptimal solution）。如：Zhu等人[131]提出的bMCL方法，迭代的利用Boosting来完成多类以及单类的物体发现；本人的ACCV 12文章[137]中嵌套的使用RPCA方法来寻找存在于同一个低秩空间的相似物体。采用凸优化来

完成这一任务能够取得全局最优解，相对迭代的方法更有吸引力。最近Lerman等人[138]提出了REAPER方法去从具有大量的离群点的数据中学习到数据模型，相对RPCA类似的方法在处理高斯离群点的情况下具有优势，但该方法对于内点样本上的噪声并不能有效的处理。本章中提出的算法既能够采用凸优化在存在大量离群点的情况下学习到数据模型，同时也能够处理内点样本上的噪声。

从物体发现的这个研究问题来讲，最近出现了大量的研究工作。在文献[139]中，每一幅图像都被视作为一个局部特征的集合，采用LDA[140]可以发现其中的一个共同特征。其它的一些系统[141,142]在利用图像中各种线索建立起来的关系矩阵上进行聚类来完成物体发现。尽管目前现有的方法在标准数据集上能够取得非常令人欣喜的物体发现结果，但它们还是存在一些局限性：（1）对于物体发现这一问题缺少一个清晰的数学表达，这导致了它们并不能在性能上得到保证，提供物体发现是否能够成功的边界条件；（2）一些系统比较拼凑，利用了各种的图像特征、分类器，表现出严重的拼凑感；（3）目前方法均严重依赖于区分性模型，泛化能力有限。

不同于之前的方法，本章从一个不同的角度去探索目标发现这一问题——利用一个简单的生成型模型来建模物体的子空间，为推断物体标记和学习物体模型提供一个新颖的统一的目标方程，并采用低秩优化技术快速的求解。不同于那些类似于EM的方法，本章中提出的方法对于初始化不敏感，能够处理物体样本上出现的错误。本章方法沿袭了RPCA系列的方法的优点，二者均是通过凸优化求解，都表现出了很强的鲁棒性和高效性。大量的仿真和实际图像数据上的实验充分体现了本章方法的优势。

4.2 子空间发现的数学框架

给定K个包含示例的包，第k,（$k \in [1,\ldots,K]$）包中有n_k个示例，这样所有包中所包含的总的示例的个数为$N = n_1 + \cdots + n_K$。每一个示例为一个d维的向量$x_i^{(k)} \in \Re^d$。另外，我们定义$X^{(k)} = \left[x_1^{(k)},\ldots,x_{n_k}^{(k)}\right] \in \Re^{d\times n_k}$。我们默认每一个包中至少存在一个所要寻找的共同物体，其它的非共同物体之间不相关。每一个示例$x_i^{(k)}$对应于一个二值的标记$z_i^{(k)} \in \{0,1\}$。$z_i^k = 1$表示x_i^k是我们需要发现的共同物体.类似的，我们定义$Z^{(k)} = \left[z_1^{(k)},\ldots,z_{n_k}^{(k)}\right] \in \{0,1\}^{n_k}$和$Z = \left[Z^{(1)},\ldots,Z^{(K)}\right]$。我们假设每一个包中至少含有一个共同物体。因此我们可以得到$\bigvee_{i=1}^{n_k} z_i^{(k)} = 1, \forall k \in [K]$，其中$\bigvee$是一个操作符且$[K] = \{1,2,\ldots,K\}$是一个包含小于或等于$K$的正整数的集合。

通常，对应于共同物体的不同示例在外观上十分相似，那么它们的特征向量高度的相关。因此，我们可以假设这些对应于共同物体的示例存在于一个低维的子空间中$\Omega \subset \Re^d$。这个假设可以从真实数据的实验中得到证实。图4-4 (c)显示了共同物体和随机选取的图像块的特征值的分布。即使采用RPCA对随机选取的图像块进行

去噪，去噪之后的特征值分布依然比共同物体的特征分布要分布。

尽管如此，因为自然图像中存在很多的干扰因素，例如物体的形变、光照变化、局部遮挡等，观测到的共同物体已经不处于一个低维空间中了。我们可以将这些干扰因素建模为示例上的稀疏错误。这样，我们可以将每个示例表示为$x = a + e$，其中$a \in \Omega$且e是一个稀疏向量。

从给定的K个包中$X = \left[X^{(1)}, \ldots, X^{(K)}\right]$, 本章的目的是从每一个包中选择一个或者多个示例并且使得选择的示例能够组成一个低秩矩阵A和稀疏错误矩阵E。对于这个问题，我们可以给出如下数学表达式：

$$\begin{aligned} &\min_{A,E,Z} \operatorname{rank}(A) + \gamma \|E\|_0 \\ &\text{s.t.} \quad X\operatorname{diag}(Z) = A + E,\ \forall k \in [K]\ \textstyle\bigvee_{i=1}^{n_k} z_i^k = 1, \end{aligned} \tag{4-1}$$

其中$\operatorname{diag}(Z)$是一个$N \times N$的包含K个$\left\{\operatorname{diag}\left(Z^{(k)}\right)\right\}$的块状对角矩阵. 为了区别于传统的“子空间学习（subspace learning）”问题，我们将这个问题成为“子空间发现（subspace discovery）”。

4.3 问题的凸优化求解

公式(4-1)中定义的问题是一个复杂的组合优化问题，其中涉及到连续变量和离散变量。在本章的应用情况中（在绝大多数情况下）d和N都会比较大，这个组合优化问题是无法求解的。最近的RPCA[143]理论证明低秩和稀疏性可以通过它们的凸近似（convex surrogate）来最小化。因此，我们可以以上目标方程中的$\operatorname{rank}(\cdot)$用核范数$\|\cdot\|_*$来近似，$\ell_0$范数用$\ell_1$范数来近似。这样，公式(4-1) 就可以采用以下形式代替:

$$\begin{aligned} &\min_{A,E,Z} \|A\|_* + \lambda \|E\|_1 \\ &\text{s.t.} \quad X\operatorname{diag}(Z) = A + E,\ \forall k \in [K]\ \textstyle\bigvee_{i=1}^{n_k} z_i^{(k)} = 1. \end{aligned} \tag{4-2}$$

可以发现，以上目标方程是一个凸优化问题，但是限制条件中二值变量$z_i^{(k)}$仍然使得这个问题很难求解。

4.3.1 一个简单的迭代求解方法

这里我们首先给出一个简单的迭代求解方法（NIM，Naive Iterative Method）。类似于EM算法，该方法的思路是变换的估计Z和最小化目标方程，即低秩部分A和稀疏错误部分E。（1）当Z已知时，公式(4-2)变成一个凸优化问题且可以采用RPCA方法[143]来直接求解。（2）一旦地址矩阵A已知，我们可以采用ℓ_1最小化来求得每一个示例到子空间的距离，即每个示例的ℓ_1错误：

$$e_i^{(k)} = \min_{w} \|Aw - x_i^{(k)}\|_1. \tag{4-3}$$

然后对于每一个包，那些ℓ_1错误小于一定阈值的示例的标记置为1，其余的示例所对应的标志置为0。我们可以不断迭代以上两个步骤直到收敛（所有示例的标记均不改变）。因为在给定数据中，大量的示例都不对应于共同物体，即离群点，这种简单的迭代求解方法对于初始化非常的敏感。因此，为了得到更好的解，我们通常进行多次随机的初始化，最终选择目标方程值最小的情况下所对应的解作为最终的解。这种思想类似于经典的RANSAC方法。假设在第k个包中有m_k个正示例，这种RANSAC式的求解方法能够成功地寻找共同物体的子空间的概率仅为$\Pi_{k=1}^{K}(\frac{m_k}{n_k})$。通常，$\forall k, m_k/n_k \leq \frac{1}{5}$，因此这种RANSAC方法成功的概率会成指数级的降低。即使正确的示例被选取了，因为A可能存在错误，以上的基于ℓ_1最小化来确定Z值的方法并不能保证得到正确的结果。然而，在谨慎的选择初始化方法的情况下，在一些比较简单的数据集中，NIM方法可以得到相当不错的结果并作为一个基准来评估其它的一些方法。

4.3.2 通过松弛Z来求解

在之前，变量Z是一个二值的$\{0,1\}$。这里，我们将它松弛到实数域$\Re$中。另外，限制$\bigvee_{i=1}^{n_k} z_i^{(k)} = 1$能够松弛为一个连续型的约束$\mathbf{1}^T Z^{(k)} = 1$，这是一个线性约束。因此，本章的优化问题变为

$$\begin{aligned} &\min_{A,E,Z} \|A\|_* + \lambda\|E\|_1, \\ &\text{s.t.} \quad X\operatorname{diag}(Z) = A + E,\ \forall k \in [K], \mathbf{1}^T Z^{(k)} = 1. \end{aligned} \tag{4-4}$$

尽管我们不去显性的要求Z是非负的，但事实上这个问题的最优解总能保证$Z^* \geq 0$。对此，我们将给出定理证明。其中的原因在于核范数和ℓ_1的优良特性。由于我们不需要显性约束Z非负，我们可以节省N非等式约束，可以显著的提高求解算法的效率。首先我们来看以下引理:

引理 4.1: 给定一个矩阵$Q \in \Re^{m\times n}$，如果$\tilde{Q}$是Q的某些列被缩小（乘上了一些$\alpha \in (0,1)$），那么$\|\tilde{Q}\|_* < \|Q\|_*$。

证明: 不失一般性，我们假设Q的最后一列q_n被缩小了。Q可以表示为$Q = [Q_{n-1},\ q_n]$，并且$Q' = [Q_{n-1},\ 0]$表示Q的最后一列被置为0。Q'的特征值为Q_{n-1}的特征值和0的交集。让$t = \min\{m,n\}$。根据文献[144] 的定理7.3.9，$\sigma_1(Q) \geq \sigma_1(Q') \geq \sigma_2(Q) \geq \sigma_2(Q') \geq \ldots \geq \sigma_t(Q) \geq \sigma_t(Q') \geq 0$。因此，很自然的，只有$\sigma_i(Q) = \sigma_i(Q'), \forall i \in [t]$成立，$\|Q\|_* \geq \|Q'\|_*$中的等号成立。因为$\|Q\|_F^2 = \sum_i \sigma_i(Q)^2 > \|Q'\|_F^2 = \sum_i \sigma_i(Q')^2$，所以这种情况是有可能的。因此，我们可以得到$\|Q\|_* > \|Q'\|_*$。

注意到$\tilde{Q}=\alpha Q+(1-\alpha)Q'$且核范数$\|\cdot\|_*$是凸的，应用Jensen不等式，我们可以得到

$$\|\tilde{Q}\|_*\leq\alpha\|Q\|_*+(1-\alpha)\|Q'\|_*<\alpha\|Q\|_*+(1-\alpha)\|Q\|_*=\|Q\|_* \tag{4-5}$$

至此引理4.1证明完毕。

定理 4.1： 假如X中任何一列不全为0，那么公式(4-4)的最优解Z^*始终非负。

证明： 假设给定一组最优解(A,E,Z)，其中Z中存在负数。让我们采用如下方式构造一个新的三元组：

$$\begin{aligned}\hat{Z}^{(k)}&=\tfrac{1}{1^T|Z^{(k)}|}|Z^{(k)}|,\\ \hat{A}^{(k)}&=\tfrac{1}{1^T|Z^{(k)}|}A^{(k)}\mathrm{diag}\left(\mathrm{sign}\left(Z^{(k)}\right)\right),\\ \hat{E}^{(k)}&=\tfrac{1}{1^T|Z^{(k)}|}E^{(k)}\mathrm{diag}\left(\mathrm{sign}\left(Z^{(k)}\right)\right)\end{aligned} \tag{4-6}$$

因为$X\mathrm{diag}\,(Z)=A+E$，显然$X\mathrm{diag}\left(\hat{Z}\right)=\hat{A}+\hat{E}$，因此，$(\hat{A},\hat{E},\hat{Z})$也是一组可行的解，而且$\hat{Z}$是非负的。我们将采用反证法的思路，推导出$\|\hat{A}\|_*+\lambda\|\hat{E}\|_1<\|A\|_*+\lambda\|E\|_1$，它与$(A,E,Z)$是矛盾的。

注意到将矩阵的任何一列的符号取反不会改变这个矩阵的特征值，因此这个矩阵的核范数不会改变（$W=U\Sigma V^*$, $\mathrm{diag}\,(\pm1,\ldots,\pm1)\,V$仍然是一个正交矩阵）。那么，我们构造另外一个矩阵$A'^{(k)}=A^{(k)}\mathrm{diag}\left(\mathrm{sign}\left(Z^{(k)}\right)\right)$，它的核范数不变$\|A'\|_*=\|A\|_*$。类似的，我们构造$E'$，$\|E'\|_1=\|E\|_1$。

$\hat{A}$ 和$\hat{E}$是A' 和E'按照列变小的版本。下面我们将做出解释：对于第k个包，$1^TZ^{(k)}=1$，假如$Z^{(k)}$中任何一项为负数，则$1^T|Z^{(k)}|>1$，否则$1^T|Z^{(k)}|=1$。因此，那些$Z^{(k)}$中负数所对应的A'和E'的列被缩小了$\alpha^k\in(0,1)$倍。接下来，根据引理4.1可以得出缩小矩阵的任何一列会降低该矩阵的核范数。

$\hat{A}$可以被看作A的多个不同的列被缩小的版本，而且，每一个列的缩小都会导致核范数的减小。同样，缩小列的值也会减小稀疏错误E的ℓ_1范数。这表明$\|\hat{A}\|_*+\lambda\|\hat{E}\|_1<\|A\|_*+\lambda\|E\|_1$——这与之前的假设$(A,E,Z)$是最优解的假设是矛盾的。

至此定理4.1证明完毕。

4.3.3 基于交替方向乘子法的求解方法

我们应用交替方向乘子法（ADMM，Alternating Direction Method of Multipliers）来对公式(4-4)的问题进行求解。首先，我们写下拉格朗日参数方程：

$$\begin{aligned} &L(A,E,Z,Y_0,Y_1,\ldots,Y_K) \doteq \|A\|_* + \lambda\|E\|_1 \\ &+\langle Y_0, X\mathrm{diag}\,(Z) - A - E\rangle + \tfrac{\mu}{2}\|X\mathrm{diag}\,(Z) - A - E\|_F^2 \\ &+\textstyle\sum_{k=1}^{K}\left(\langle Y_k, \mathbf{1}^T Z^{(k)} - 1\rangle + \tfrac{\mu}{2}\|\mathbf{1}^T Z^{(k)} - 1\|_F^2\right). \end{aligned} \tag{4-7}$$

因为上式中存在三个变量，直接应用增强拉格朗日乘子法解决上面的问题是不可行的，所以我们采用一种在文献[113,133]中提出的策略。具体做法是交替地去优化这三个变量，如在第t次迭代中：

$$\begin{cases} A_{t+1} = \underset{A}{\mathrm{argmin}}\, L(A, E_t, Z_t, Y_t, \mu_t) = \\ \quad \underset{A}{\mathrm{argmin}}\, \|A\|_* + \dfrac{\mu_t}{2}\left\|X\mathrm{diag}(Z_t) - A - E_t + \dfrac{Y_{0,t}}{\mu_t}\right\|_F^2, \\ E_{t+1} = \underset{E}{\mathrm{argmin}}\, L(A_{t+1}, E, Z_t, Y_t, \mu_t) = \\ \quad \underset{E}{\mathrm{argmin}}\, \|E\|_1 + \dfrac{\mu_t}{2}\left\|X\mathrm{diag}(Z_t) - A_{t+1} - E + \dfrac{Y_{0,t}}{\mu_t}\right\|_F^2, \\ Z_{t+1} = \underset{Z}{\mathrm{argmin}}\, L(A_{t+1}, E_{t+1}, Z, Y_t, \mu_t) = \\ \quad \underset{Z}{\mathrm{argmin}}\left\|X\mathrm{diag}(Z) - A_{t+1} - E_{t+1} + \dfrac{Y_{0,t}}{\mu_t}\right\|_F^2 + \\ \qquad\qquad \ldots \sum_{k=1}^{K}\left\|\mathbf{1}^T Z^{(k)} - 1 + \dfrac{Y_{k,t}}{\mu_t}\right\|_F^2. \end{cases} \tag{4-8}$$

幸运的是，以上三个最小化问题都能够得到闭合解。细节的推导过程在如下段落中给出。

采用$\mathcal{S}_\epsilon(\cdot)$表示一个收缩操作符定义如下。

$$\mathcal{S}_\epsilon(x) = \begin{cases} x - \varepsilon, & \text{if} \quad x > \varepsilon, \\ x + \varepsilon, & \text{if} \quad x < -\varepsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{4-9}$$

如果svd分解表示为$X\mathrm{diag}\,(Z_t) - E_t + \frac{Y_{0,t}}{\mu_t} = U\Sigma V^*$，$A_{t+1}$的最优值为$A_{t+1} = U\mathcal{S}_{\frac{1}{\mu}}(\Sigma)V^*$。对于$E_{t+1}$，最优值为$\mathcal{S}_{\frac{\lambda}{\mu_t}}\left(X\mathrm{diag}\,(Z_t) - A_{t+1} + \frac{Y_{0,t}}{\mu}\right)$。对于$Z$，我们可以将原始的优化问题分解为$K$个针对$Z^{(k)}$的优化问题。每一个子优化问题是一个典

型的最小二乘法优化问题。

$$Z_{t+1}^{(k)} = \underset{Z^{(k)}}{\operatorname{argmin}} \left\| X^{(k)} \operatorname{diag}\left(Z^{(k)}\right) - A_{t+1}^{(k)} - E_{t+1}^{(k)} + \frac{Y_{0,t}^{(k)}}{\mu_t} \right\|_F^2 \\ \ldots + \left\| \mathbf{1}^T Z^{(k)} - 1 + \frac{Y_{k,t}}{\mu_t} \right\|_F^2 \tag{4-10}$$

为了简介的表示，我们采用$P^{(k)} = A_{t+1}^{(k)} + E_{t+1}^{(k)} - \mu_t^{-1} Y_{0,t}^{(k)} \in \Re^{d \times n_k}$，并且我们将$P^{(k)}$第$i$列记为$P_i^{(k)}$，$Q^{(k)} = 1 - \mu_t^{-1} Y_{k,t} \in \Re^1$。另外，我们定义$X_R^{(k)} = \begin{bmatrix} x_1^{(k)} & & \\ & \ddots & \\ & & x_{n_k}^{(k)} \end{bmatrix}$ and $P_R^{(k)} = \operatorname{vec}\left(P^{(k)}\right)$。这样公式(4-10) 重写为

$$\begin{aligned} Z_{t+1}^{(k)} =& \underset{Z^{(k)}}{\operatorname{argmin}} \left\| X_R^{(k)} Z^{(k)} - P_R^{(k)} \right\|_F^2 + \left\| \mathbf{1}^T Z^{(k)} - Q^{(k)} \right\|_F^2 \\ =& \left\| \begin{bmatrix} X_R^{(k)} \\ \mathbf{1}^T \end{bmatrix} Z_{t+1}^{(k)} - \begin{bmatrix} P_R^{(k)} \\ Q^{(k)} \end{bmatrix} \right\|_F^2 \end{aligned} \tag{4-11}$$

直接应用标准的最小二乘法需要计算$X_R^{(k)} \in \Re^{(dn_k+1) \times n_k}$的伪逆，计算这样一个高维矩阵的伪逆（pseudo-inverse）是一个计算量很大的问题。因此我们采用一个技巧使得解决这个最小二乘法仅需要计算一个$\Re^{n_k \times n_k}$矩阵的伪逆。具体过程如下：

$$\begin{aligned} Z_{t+1}^{(k)} =& \left(\begin{bmatrix} X_R^{(k)T} & \mathbf{1} \end{bmatrix} \begin{bmatrix} X_R^{(k)} \\ \mathbf{1}^T \end{bmatrix} \right)^\dagger \begin{bmatrix} X_R^{(k)T} & \mathbf{1} \end{bmatrix} \begin{bmatrix} P_R^{(k)} \\ Q \end{bmatrix} \\ =& ((X_R^{(k)T}) X_R^{(k)} + \mathbf{1} \cdot \mathbf{1}^T)^\dagger (X_R^{(k)T} P_R^{(k)} + \mathbf{1} \cdot Q^{(k)}) \\ =& \begin{bmatrix} (x_1^{(k)})^T x_1^{(k)} + 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & (x_{n_k}^{(k)})^T x_{n_k}^{(k)} + 1 \end{bmatrix}^\dagger \\ & \ldots \begin{bmatrix} (x_1^{(k)})^T P_1 + Q^{(k)} \\ \vdots \\ (x_{n_k}^{(k)})^T P_{n_k} + Q^{(k)} \end{bmatrix} \end{aligned} \tag{4-12}$$

在$A, E,$ 和Z都更新之后，我们仅需要在对偶变量Y_t进行梯度上升：

$$Y_{0,t+1} = Y_{0,t} + \mu_t \left(X \operatorname{diag}(Z_{t+1}) - A_{t+1} - E_{t+1} \right),$$

$$Y_{k,t+1} = Y_{k,t} + \mu_t \left(\mathbf{1}^T Z_{t+1}^{(k)} - 1 \right).$$

另外，μ通过$\mu_{k+1}=\rho\mu_k, \rho>1$来更新。

完整的算法在一面的算法4-1中描述。

算法 4-1 基于交替方向乘子法的鲁棒的子空间发现算法

Input: 包X 和参数λ

$Z_0=0, Y_0=0; \forall k\in[K], Y_{k,0}^{(k)}=0; E_0=0; \mu_0>0; \rho>1; t=0$

while 未收敛 **do**

// 第4-5行解决$A_{t+1}=\mathrm{argmin}_A\, L(A, E_t, Z_t, Y_t, \mu_t)$.

$[U,\Sigma,V^*]=\mathrm{svd}(X\mathrm{diag}\,(Z_t)-E_t+\frac{Y_{0,t}}{\mu_t})$;

$A_{t+1}=U\mathcal{S}_{\frac{1}{\mu}}(\Sigma)V^*$.

// 第7 行解决$E_{t+1}=\mathrm{argmin}_E\, L(A_{t+1}, E, Z_t, Y_t, \mu_t)$.

$E_{t+1}=\mathcal{S}_{\frac{\lambda}{\mu_t}}\left(X\mathrm{diag}\,(Z_t)-A_{t+1}+\frac{Y_{0,t}}{\mu}\right)$.

// 第9-12行解决$Z_{t+1}=\mathrm{argmin}_Z\, L(A_{t+1}, E_{t+1}, Z, Y_t, \mu_t)$.

for $k=1\to K$ **do**

得到$Z_{t+1}^{(k)}$ via 公式(4-12).

end for

$Z_{t+1}=[Z_{t+1}^{(1)},\ldots,Z_{t+1}^{(K)}]$.

// 第14-16 行更新Y_{k+1} and μ_{k+1}.

$Y_{0,t+1}=Y_{0,t}+\mu_t\,(X\mathrm{diag}(Z_{t+1})-A_{t+1}-E_{t+1})$.

$Y_{k,t+1}=Y_{k,t}+\mu_t\left(\mathbf{1}^T Z_{t+1}^{(k)}-1\right), \forall k\in[K]$.

$\mu_{t+1}=\rho\mu_t$.

$t\leftarrow t+1$.

end while

Output: 收敛后得到的(A, E, Z)

公式(4-8)中交替最小化的过程被称作是交替方向乘子法（ADMM）[132]。文献[113,133]对ADMM算法做了综述，并将其引入到低秩优化这个领域当中。ADMM算法并不能保证一定收敛至最优解。如果在目标方程中仅存在两个变量时，ADMM算法的收敛性可以得到充分的保证[132]。然而，当目标方程中的变量超过两项时，ADMM的收敛性得到了很多实际结果的验证[112]，但理论研究的结果相对较少。文献[145]对于一类具有三个未知变量的方程给出了收敛性的严格证明（应用在有噪声的主成分发掘问题上）。但是，文献[145]中的情况与本文中的公式(4-8)中的问题不同，无法应用其理论结果在本章的问题上。通用的ADMM算法的收敛性仍然是一个开放问题。尽管如此，目前在这方面不断有新的研究出现。对于本文中的ADMM问题，文献[146]的思路是值得借鉴的，例如，我们可以将变量E和Z放在一起，然后应用近端的ADMM算法（Proximal ADMM）。这种近端的ADMM算法可以保证收敛性。然而，在实际应用中，由于我们将A，E和Z分开，我们可以得到

比近端的ADMM算法更快的收敛速度。根据我们的经验，本文提出的这种ADMM算法拥有非常快的速度，满足实际的应用需求。

4.4 仿真和实验

在本节中，我们将在人工生成的数据上进行仿真以及真实的图像数据来进行实验，通过各种不同的任务来验证本章提出的算法的有效性。我们将章节4.3.1中的方法记为NIM方法，并将松弛后的算法记为ADMM方法。在所有的仿真和实验中，我们将参数λ设置为$\lambda = 1/\sqrt{d}$，其中d是示例的维度。

4.4.1 子空间学习仿真

为了去研究我们提出的ADMM方法在子空间发现这个任务上的能力，在这个仿真中，我们人工生成50个包：在每个包中有10个示例，包含1个正的示例和9个负的示例。示例的特征的维度为$d = 500$。在生成正示例时，我们首先随机生成r个$d \times 1$维向量作为基向量，基向量的每一维属于i.i.d.标准高斯，再将基向量用随机系数线性组合来生成正的示例。负示例是独立的随机的生成的满足i.i.d.标准分布的$d \times 1$向量。对于每一个示例，无论它是正的示例还是负的示例，我们将它进行归一化，使得它的ℓ_2范数为1。最后，大的稀疏的错误被加到了所有的示例上。错误的稀疏比率记为s，值均匀的分布在区间$[-1, 1]$中。

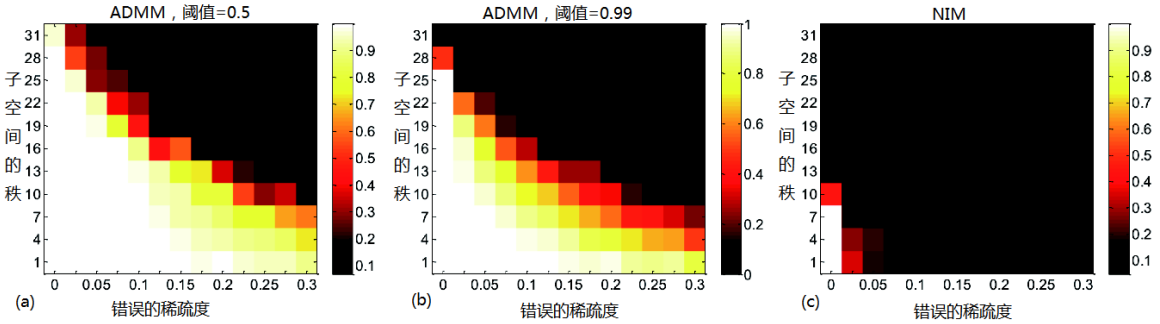


图 4-2 当错误的稀疏度和子空间的秩变化时，NIM算法和ADMM算法在不同的阈值下恢复的指示的准确度。(a)阈值$\tau = 0.5$情况下的ADMM方法。(b)阈值$\tau = 0.99$情况下的ADMM方法。(c)NIM方法。

我们研究了当子空间的秩r和错误的稀疏度s变化对于子空间发现的精度的影响。r 变化的范围是从1到31；s 变化的范围是从0到0.3。对于每一次仿真，我们将真实的指示向量记为Z^*，恢复出来的指示向量记为$\hat{Z}$，集合I^*中包含Z^*中值为1的索引，且集合$\hat{I}$中包含$\hat{Z}$中值大于某一阈值$\tau \in [0, 1]$的索引。接下来，我们定义恢复的指示向量的准确度：$\texttt{accuracy} = \frac{\#(I^* \cap \hat{I})}{\#(I^*)}$。固定一组$r$和$s$，我们随机的生成数据5次，用

算法进行测试并计算指示向量的准确度。图4-2中显示了不同阈值下的ADMM算法和NIM算法的5次测试的平均精度。在图4-2(a)中，$\tau = 0.5$，0.5是一个很公平的阈值，因为每个包中恢复的指示中最多只有一个的值大于0.5。在图4-2(b)中，$\tau = 0.99$，这是一个十分严格的阈值，在这个显示了章节4.3.2中松弛了变量Z是否还能保持变量Z的精确性。在图4-2(c)中，NIM方法得到的解释离散的，所以不需要设置阈值来将结果二值化。从图4-2中的结果来分析，NIM方法只能够在没有错误，且子空间的秩很小的情况下工作，而ADMM方法可以工作的区域显然更大。对比图4-2(a)和(b)中的结果，我们可以发现当子空间的维度越低、示例上的错误越大就越有利于ADMM算法精确的恢复指示向量。

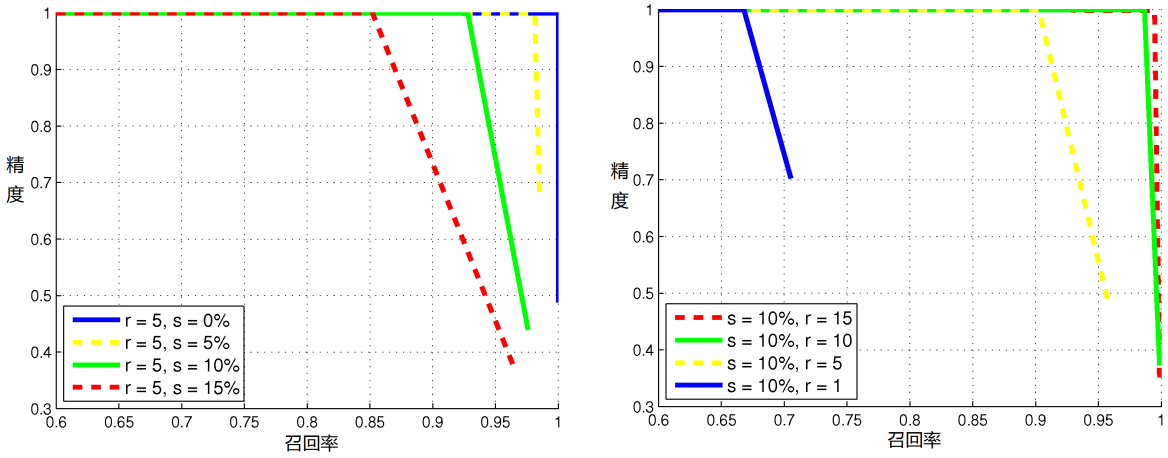


图 4-3 ADMM方法恢复出的指示向量的精度-召回率（Precision-recall）曲线。左图: 子空间的秩固定为$r = 5$; 错误的稀疏度$s = 0\%, 5\%, 10\%, 15\%$. 右图: 错误的稀疏度固定为$s = 10\%$; 子空间的秩$r = 1, 5, 10, 15$。

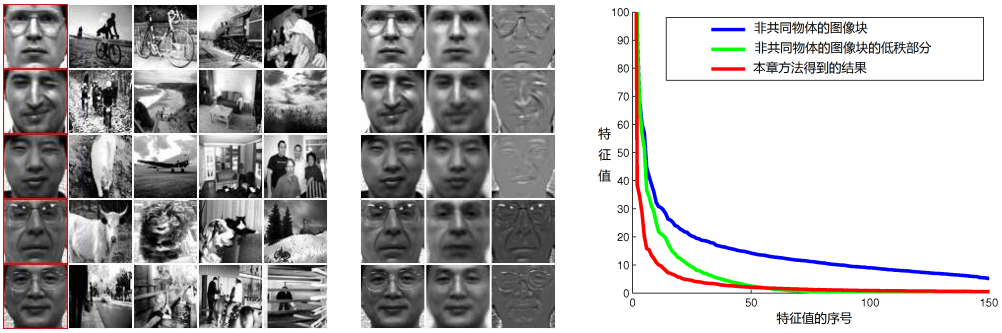
一个包中含有多个正示例的情况： 以上的仿真都是建立在一个包中仅有一个正的示例的情况下的。接下来，我们将研究在一个包中含有多个示例的情况下ADMM算法的性能。在每个包中，我们放置3个正的示例，同样，这3个示例也是随机从子空间中生成并添加幅度大的稀疏噪声的。这样，每个包中3个正的示例都是不同的。对于不同的r 和s值，我们同样5次随机生成数据并运行ADMM。我们采用精度-召回率（precision-recall）曲线来描述恢复出的指示向量的质量。给定一个阈值τ，精度和召回率采用以下公式来计算：$\texttt{precision} = \frac{\#(I^* \cap \hat{I})}{\#(\hat{I})}$ 和$\texttt{recall} = \frac{\#(I^* \cap \hat{I})}{\#(\texttt{all positive instances})}$。如图4-3(a)所示，ADMM的性能随着错误的稀疏程度的增加而提升，当示例上没有错误的情况下，ADMM算法能够完美地寻找到所以的正示例。图4-3(b)显示子空间的维度较高有利于发现更多的正示例——这个现象是有道理的：因为如果正示例的个数小于自控将的维度，我们无法争取的得到子空间的维度。

当子空间的维度为15，错误的稀疏度为10%时，ADMM算法能够保证在100%的召回率情况下得到99%的正示例。然而，在目前我们针对子空间发现的问题定义（公式(4-4)）下，我们并没有一个线性的机制去要求在任何情况下都能寻找到所有的正示例。

4.4.2 随机图像块中的对齐的人脸发现

在这个仿真中，目的是在大量的随机采样得到的图像块中寻找对齐好的人脸图像。人脸图像全部来自于Yale人脸数据集[147]，这个数据集中含有来自于15个人的165张正脸图像。其它的图像块随机的从PASCAL数据集[148]中采样得到。我们按照如下的方法来设计本次仿真中的包和示例：165张正脸图像被分别放置在165个包中，每个包中除了一张人脸图像外还含有9个从PASCAL数据集中随机采样得到的图像块。每个人脸/图像块的大小都归一化为64×64像素，然后展开成一个4096维的向量作为特征。图4-4(a)中展示了一些包中的图像块和人脸。

为了评估人脸发现这个任务的性能，我们选择每个包中恢复出来的指示值最大的示例，并计算这些被选择的示例中Yale人脸的比率作为准确度。由于负示例的选择存在随机性，我们将这个仿真跑5次并计算平均的准确度。ADMM算法和NIM算法（随机初始化）的平均准确度分别为99.5±0.5% 和77.8±3.5%。一些采用ADMM算法选择得到的人脸在图4-4(b)中显示。如图所示，人脸上的一些表情、眼睛等噪声被去除了，这使得到的子空间处在一个较低的维度上。



(a) (b) (c)

图 4-4 (a): 每一行一个包中采样的一些图像。(b): 采用ADMM算法得到的人脸发现的结果。第一列显示原始的图像；第二列和第三列分别显示恢复的低秩部分和错误部分。(c): 采用ADMM算法得到的低秩部分的特征值分布（红色），在非共同物体的图像应用RPCA[143]之后得到的低秩部分的特征值分布（绿色），非共同物体的图像的原始的特征值分布（蓝色）。

4.4.3 真实场景中的物体发现

近年来，物体发现逐渐成为了计算机视觉中的一个主要问题，通过物体发现这个工具可以显著地减少图像标注工作，同时物体发现也是一个十分具有挑战性的任务。在这个任务中，我们给定的是一堆图像，其中每张图像都含有一个或者多个共同物体。不同于之前的有监督的物体检测设定，在物体发现这个任务中，我们没有任何已经标定物体位置的图像。不同于在人工生成的数据上进行仿真，在真实的场景中，物体的外观有着十分显著的变化，因此，我们需要采用一些更加鲁棒的特征描述子，如HOG特征与LBP特征。另外，由于物体在图像中的位置和大小都是未知的，从一幅图像中可以提取的示例数目可以达到数百万计，为了解决这个问题，我们采用一些现有的非监督的显著性物体检测算法[149]来减少每幅图像中的示例的个数。我们之所以可以采用HOG和LBP特征来刻画物体是因为共同物体会有着类似的纹理和形状，HOG和LBP特征擅长于刻画物体的纹理和形状，在一些有监督的物体检测方法中[150,151]它们也表现了十分优异的性能。

在这个真实场景中物体发现的实验中，我们在四个差异很大的数据集上评估了ADMM算法的性能并与业内结果最好的方法（the state-of-the-arts）进行比较。这四个数据集分别是：PASCAL 2006数据集[152]，PASCAL 2007数据集[153], FDDB（Face Detection Data Set and Benchmark）数据集[154]，和ETHZ Apple logo数据集[155]。因为使用了不同的性能评价标准，我们将分两个不同的小节来分别给出实验结果。

4.4.3.1 PASCAL 2006 和2007 数据集

PASCAL 2006 和2007 数据集是两个非常具有挑战的数据集，它们经常被用来评估有监督的物体检测和图像分类系统的性能。为了进行物体发现这个实验，我们采用标准的测试方法[156]，介绍如下：首先我们定义所谓的CorLoc度量，CorLoc度量表示的是所有被选择的示例中正确定位物体位置的比率，那么如何定义正确定位物体位置呢? 标准的做法是采用PASCAL准则（即，检测窗口和物体真实的包围盒的交集与并集的比值大于0.5）。我们从PASCAL 2006 和2007 数据集中分别取两个子集，称为PASCAL06-6×2，PASCAL06-all，PASCAL07-6×2，和PASCAL07-all。PASCAL06-6×2中含有来自于12个视角/类别的779幅图像；PASCAL06-all中含有来自于所有的33个视角的2184幅图像；PASCAL07-6×2中含有来自于12个视角/类别的463幅图像；PASCAL07-all中含有来自于所有的45个视角的2047幅图像。更多关于这个数据集的细节以及性能评估请参考原始文献[156]。

如之前所提到的，我们将每一幅图像看作是一个包，由显著性物体检测算法[149]检测到的一个图像块看作是一个示例。显著性物体检测算法[149]中的阈值参数记为τ_s，它控制了示例的个数。对于检测到的每一个图像块，即示例，我们采用标准的HOG和LBP特征来描述。对于PASCAL06-6×2 和PASCAL06-all这两个数据集，我们设置$\tau_s = 0.22$；对于PASCAL07-6×2 and PASCAL07-all这两个数据集，我

们设置$\tau_s = 0.165$。我们运行ADMM算法，并且选择每个包中指示值大的示例（图像块）作为发现的物体。在表4-1中，我们给出了ADMM算法的结果并且与其它的一些完成同样任务的算法[139,156–159]做比较。

与之前流行的一些物体发现方法[139,158,159]相比，表4-1中显示了我们的方法能够取得更好的结果。目前最好的性能由方法[156,157]给出，这两个分别使用了额外的包围盒标注信息和复杂的物体模型[150]。我们的方法采用了一个十分简单的生成型模型，并且十分的干净简单。图4-5给出了ADMM方法PASCAL-all数据集上的物体发现结果。

表 4-1 PASCAL 2006 和2007数据集上，在CorLoc准则下的物体发现结果。

方法	PASCAL06-6×2	PASCAL06-all	PASCAL07-6×2	PASCAL07-all
ESS[159]	24	21	27	14
文献[139]	28	27	22	14
文献[158]	45	34	33	19
ADMM（本章方法）	57	43	40	27
文献[156]	**64**	**49**	50	28
文献[157]	N/A	N/A	**61**	**30**

4.4.3.2 FDDB子集和ETHZ Apple logo数据集

FDDB子集中包含440幅包含人脸的图像。ETHZ Apple logo数据集上包含36幅包含苹果公司标志的图像。在这两个数据集上，不同图像中的物体的外观差异非常大，而且物体的背景也十分嘈杂。在这次实验中，我们仅采用HOG作为特征，HOG特征主要刻画物体的形状特征。对应于本文中的定义，在这个实验中低秩项对应于物体的共同形状结构，稀疏错误项对应于物体上的缺失、遮挡等外观差异。我们运行ADMM算法来每一个示例的指示值，对于每一幅图像，我们将其中所有示例所对应的指示值都除以该图像中最大的指示值来进行归一化，这种归一化之后的指示值用来表示每一个示例的分数。基于这个分数，我们可以构建物体发现的性能度量。

按照PASCAL标准，如果图像块与物体真实位置的交集比上它们的并集大于0.5则认为该图像块是一个正确的检错。物体发现的性能采用两种度量来表示：1）通过不断调节阈值而得到的精度-召回率（precision-recall）曲线[148]，2）通过平均得到的不同召回率下的精度而得到的平均精度（AP，average precision）[148]。

我们将ADMM算法同四个不同的方法进行比较：基线方法显著性物体检测方法（SD，saliency detection method）[149]，一个叫做bMCL[131]基于区分型模型的物体

图 4-5 红色方框：ADMM算法在非常具有挑战的PASCAL 2007数据集上取得的物体发现的结果。绿色方框：真实的物体的位置。从上至下：物体的类别分别为：飞机、自行车、公交车、摩托车、植物和显示器。

表 4-2 以平均精度（AP）为度量，SD[149]方法，bMCL[131]方法，本章的NIM-SD方法，NIM-Rand方法，以及ADMM方法在FDDB子集上的性能比较。

方法	FDDB subset	ETHZ Apple logo
SD	0.148	0.532
bMCL	0.619	0.697
NIM-SD	0.671	0.826
NIM-Rand	0.669	0.726
ADMM（本章方法）	**0.745**	**0.836**

检测方法，采用显著性值的简单迭代方法（NIM-SD），和随机初始化的简单迭代方法（NIM-Rand）。实验中，我们调整这四个方法的参数，尽可能的使得这四个方法能够在这些数据集上能够取得更好的结果。NIM-Rand方法的平均精度是运行该方法三次而得到的平均值。四个方法与ADMM算法的结果比较在表4-2中。正如我们所发现的，ADMM方法在显著性物体检测的基础上大幅度的提高性能，并且可以超过其它比较的方法。图4-6中给出了对应的精度召回率曲线。SD方法是一个纯的自底向上的方法，没有任何先验信息。另外的四个方法采用了给定的图像当中存在一个共同的物体。bMCL方法是一个区分性方法，在一些背景比较简单的数据集，如SIVAL数据集[160]上，它能够取得业界最好的结果。然而，FDDB数据集中的情况更为复杂，对于类似于bMCL方法的区分性方法，它们很难建立一个很好的背景模型来区分背景和前景。ADMM和NIM-SD方法都能很好的应对背景嘈杂的问题，因为它们都只注重于前景物体的建模，这是一个十分好的性质。图4-7中给出了部分采用SD，bMCL，NIM-SD和ADMM方法检测到的共同物体。

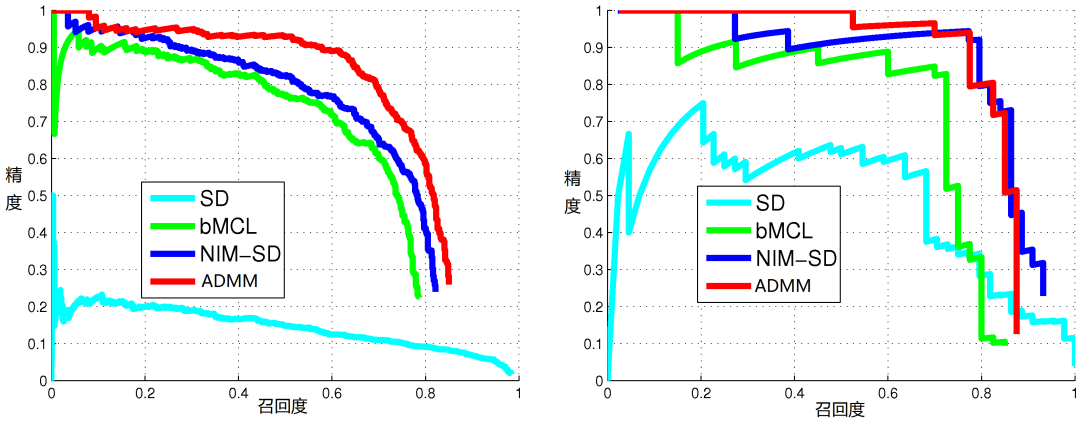


图 4-6 不同方法进行物体发现任务所得到的精度召回率曲线。青色为SD方法，绿色bMCL方法，蓝色为NIM-SD方法，另外红色为ADMM方法。左图为FDDB子集上的结果；右图为ETHZ Apple logo数据集的结果。

在这些实验中，我们发现在下列情况ADMM算法会失败：（1）所要寻找的共同物体不包含在显著性物体检测的结果之内；（2）共同物体存在比较大形变或者外观差异，导致它们不可能存在于一个共同低维子空间中。注意到，在本文中，我们更侧重于假设共同的模式存在于同一个低秩空间下的一个子空间学习问题。

4.4.4 基于示例选择的多示例学习方法

在这个实验中，我们显示了如何把我们所提出的ADMM方法应用到传统的多示例学习问题[134]当中。我们的基本想法是通过ADMM方法去鉴别所有正的包中正的示例和负的示例。鉴别出来的正示例会同所有负包中的负示例一起用来训练一个示

图 4-7 FDDB子集上的人脸发现结果。青色代表SD方法，绿色代表bMCL方法，蓝色代表NIM-SD方法，另外红色代表ADMM方法。

例级的分类器，如一个RBF核的SVM分类器，来完成多示例学习这个任务。在测试阶段，我们采用学习得到的示例级分类器来做包的分类，这是基于标准的noisy-or模型：假如一个包中存在正的示例，那么这个包是一个正包，否则，这个包是一个负包。

为了采用ADMM方法来鉴别正示例和负示例，我们采用本文中一贯采用的假设：正的示例之间是线性相关的，在去除噪声之后，它们存在于一个低维的线性空间当中。在实际中，我们将所有的正包输入到算法4-1所描述的ADMM算法中，并且得到每个示例的指示值。对于每一个包，其中所有示例的指示值都除以最大的那个指示值实现归一化。如果归一化后的指示值高于一个上限τ_u，则对应的示例会被认为是正示例；如果归一化后的指示值低于一个下限τ_l，则对应的示例会被认为是一个负示例。在我们的实验当中，我们设置$\tau_u = 0.7$且$\tau_l = 0.3$。归一化后的指示值在0.3和0.7之间的示例将被忽略且不用于训练分类器。我们选择的分类器为RBF核的SVM分类器，我们采用了LibSVM[161]中的实现。

我们将上述基于ADMM的多示例学习方法在五个标准的多示例学习测试集上进行测试，它们分别是*Musk1*，*Musk2*，*Elephant*，*Fox*和*Tiger*数据集。关于这五个数据集的具体描述可以参考之前的文献[51,162]。我们将我们所提出的方法同一些经典方法以及目前最新的一些方法进行比较，包括MI-SVM和mi-SVM方法[51]、MILES方法[163]，EM-DD方法[164]，PPMM Kernel方法[165]，MIGraph and miGraph方法[166]和MI-CRF方法[167]。在测量实验结果时，我们采用标准做法，对数据进行10份的交叉验证，并在表4-3给出了10次测试的平均准确率和标准差。由于之前的一些方法没有记录标准差，将不在表中给出。五个数据集上的平均精度的平均值在最右侧的列中给出。对于每一个比较项，我们采用黑体来显示出最好的方法的结果。

正如我们从表4-3中可以看出的那样，MIGraph和miGraph方法可以得到最好的结果，原因在于它们能够利用示例之间的图结构关系，而其它的方法假设各个示

例之间是相互独立的。我们侧重于将ADMM方法同mi-SVM方法进行比较，因为mi-SVM方法也是基于示例选择机制的。mi-SVM的机制是在MIL的约束下通过迭代的最大化正示例和负示例之间的分界面来选择正示例的，这是一个非凸问题，并且无法保证得到一个局部最优解。而这里，我们提出的ADMM方法可以采用一个凸优化的形式得到正示例所在的子空间，获得比mi-SVM算法更好的结果。

表 4-3 在五个多示例学习标准数据集上MI-SVM和mi-SVM方法[51], MILES方法[163], EM-DD方法[164], PPMM Kernel方法[165], MIGraph和miGraph方法[166], MI-CRF方法[167],以及本章提出的方法的各类准确度和平均准确度(%)的比较。

数据集	*Musk1*	*Musk2*	*Elephant*	*Fox*	*Tiger*	Average
MI-SVM	77.9	84.3	81.4	59.4	84.0	77.4
mi-SVM	87.4	83.6	82.0	58.2	78.9	78.0
MILES	86.3	87.7	-	-	-	-
EM-DD	84.8	84.9	78.3	56.1	72.1	75.2
PPMM Kernel	**95.6**	81.2	82.4	60.3	80.2	79.9
MI-CRF	87.0	78.4	85.0	65.0	79.5	79.0
ADMM(本章方法)	89.9±0.7	85.0±1.6	79.6±0.9	**65.4±1.2**	81.5±1.0	80.3
MIGraph	90.0±3.8	90.0±2.7	85.1±2.8	61.2±1.7	81.9±1.5	81.6
miGraph	88.9±3.3	**90.3±2.6**	**86.8±0.7**	61.6±2.8	**86.0±1.6**	**82.7**

4.4.5 运行效率

本章中的ADMM算法中计算资源主要消耗在其中的SVD步骤中，对于一个$m \times n, (m \leq n)$的矩阵，SVD的时间复杂度为$\mathrm{O}\left(m^2 n\right)$。那么本章算法的时间复杂度为$\mathrm{O}\left(m^2 nk\right)$，$k$为ADMM算法收敛所需要的迭代次数，一般为100左右。

4.5 本章小结

在本章中，我们提出了一种新颖的数学框架来进行弱监督的鲁棒的子空间发现。我们将一个高维的组合优化问题松弛为一个凸优化问题，并采用交替迭代乘子法对该问题进行高效的求解。不同其它的一些基于区分型模型的物体模型学习方法，我们所提出的方法可以从输入数据中寻找到感兴趣物体所在的子空间。在不同的应用以及不同的数据集上，我们充分的展示了本章所提出的方法的优势。对于后续研究，本章的研究内容的意义在于：利用多示例学习的约束以及子空间的低秩假设，我们可以在存在大量离群点的情况下发现子空间并寻找到子空间中包含的样本，这使得一系列子空间学习方法能够得到更广阔的应用，特别是那些基于RPCA方法的应用。

5 最大化间隔的多示例学习

高效的图像表示是物体识别的关键，在近年来的一些成功的图像表示当中都显性或隐性地存在着码本（字典）学习的过程，如词袋模型[168]、稀疏表示[169]、深度神经网络[170]等，码本的学习自然成为了物体识别中的核心问题。本章介绍一种基于最大间隔多示例学习的码本学习方法（MMDL，Max-margin Multi-instance Dictionary Learning），MMDL方法在一种弱监督的情况下，通过区分性学习挖掘出有助于物体识别的语义，并将这些有语义的图像特征用于图像表示的码本。MMDL将码本的学习与图像中语义的挖掘融合在一个框架中，提出了一个统一的学习框架。在众多标准的图像分类测试集上，本章中所提出的图像表示方法取得了优异的图像分类精度，且具有特征维度低、识别速度快的优点。

5.1 研究现状

在计算机视觉的应用中，一个很典型的情况是给定一个码本，对于输入图像中的每一个图像块的特征根据这个码本计算出一个分布，这个过程中计算分布的过程被称为是特征编码，之后在编码后的特征上进行简单的统计之后就可以得到图像的表示。采用码本建立图像表示有如下好处：

- 对于图像给出了一个直观的表示，这个表示往往是一个简单的向量。码本中的每一个单位对应于不同的语义（可能是底层的、中层的或是高层的），基于码本的图像表示可以直接用来计算图形间的相似度（或距离），不同于原始的图像像素或是一堆图像中局部特征的集合。
- 对图像实现了降维。特征提取之前的图像存在无限高的维度当中，特征编码将图像的维度降到同码本尺寸的维度，为后续识别工作降低了复杂度。
- 通过调整码本的尺度，可以实现对图像进行层次化的表达（hierarchical representations）。
- 在基于码本的图像表示中可以很轻易的描述图像特征之间的空间关系。

最直观的学习图像码本的方法是采用无监督的聚类方法，如经典的k-means算法[171]，以及基于k-means的一些变体[172,173]。但是直接的进行聚类具有对初始化、码本尺寸、聚类度量敏感等问题。从另一个方面讲，人们发现使用已知的监督信息可以学习到更有区分型的码本[12,74,174–177]，区分型码本学习方法的本质选出图像中具有区分性的特征。此外，还有一些方法[178–180]尝试去分析图像中内在的结构关系，并将这些结果关系附加在图像码本当中。但这些方法并不是通用的码本学习方法。最近流行的图像属性[53,181,182]（Attributes）、小姿势[183]（Poselets）以及物体组[184]（Object Bank）等方法都属于基于码本的图像表示方法，这些方法展示了基于码本的图像表

示这个方向的广阔前景。但这些方法在训练的时候都需要非常强的监督信息，即标注出对应训练样本，这样人工标注的开销很大。本章中研究在一种弱监督的情况（即只有图像标记给定）下的图像码本学习，如图5-1所示。具体来讲，本章中将这个弱监督学习问题定义为多示例学习（MIL）问题：可以将一幅图像看作MIL中的一个包，将图像中的一个图像块当做为一个看作MIL中的一个示例。通过MIL正好可以找到有区分性的示例，再采用聚类方法得到码本。

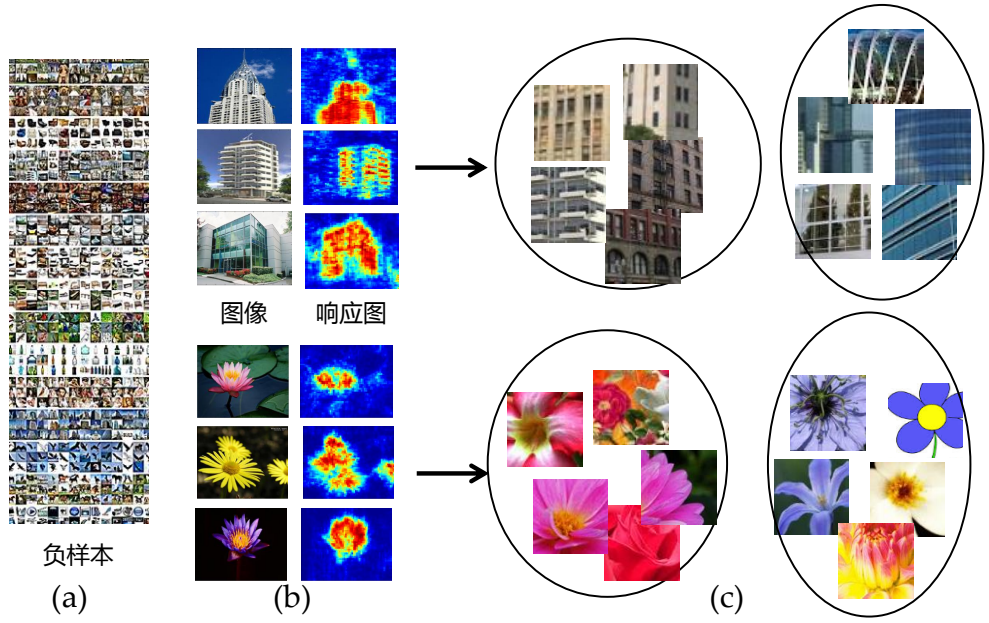


图 5-1 基于弱监督学习的图像码本学习。（a）显示了图像负样本，它们的类别不同于（b）中的图像。在（b）中，对于每一个图像类别，可以利用负样本作为支撑，采用多示例学习找到其中的区别于负样本的区域，如（b）中第二列的响应图所示，红色区域对应响应高的区域。（c）中利用（b）中寻找到响应高的区域聚类得到码本。

从MIL的角度来讲，本章的方法跟多成分学习[185]（MCL，Multiple Component Learning）和多中心的多示例学习[186]（MCIL，Multiple Clustered Instance Learning）类似，但这些方法并不关注于图像中码本学习这一问题。另外，采用区分型方法完成MIL工作的还有M^3MIML[187]和M^3IC[188]方法。但本章的多示例学习方法同它们有着显著的区别：M^3MIML和M^3IC的目的在于最大化包之间的间隔，而本章中的方法在于最大化示例间的间隔。

5.2 最大化间隔的多示例学习的定义

5.2.1 数学表示和研究动机

在介绍本章新提出的MIL算法之前，首先简单地回顾MIL的通用的数学表示。在MIL中，给定的是一系列包（Bag）$X = \{X_1, \ldots, X_n\}$，每一个包包括一系列示例（Instance）$X_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im}\}$，每一个示例是一个$d$维的向量$\mathbf{x}_{ij} \in \mathbf{R}^{d\times 1}$。另外，每一个包都有一个相应的标记$Y_i \in \{0, 1\}$——表示一个包是负的还是正的；每一个示例也都有一个示例标记$y_{ij} \in \{0, 1\}$。包的标记和示例的标记之间的是如下关系：如果$Y_i = 0$，那么对所有$j \in [1, \ldots, m]$都有$y_{ij} = 0$，即包中所有示例都是负样本。如果$Y_i = 1$，那么至少有一个$\mathbf{x}_{ij} \in X_i$示例是正样本。

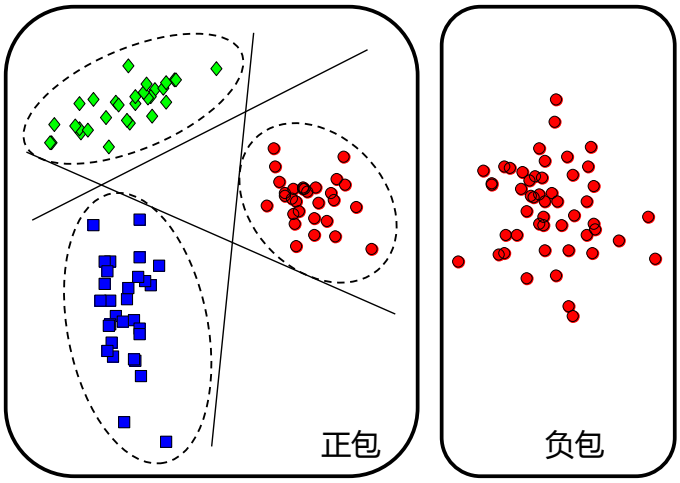


图 5-2 最大化间隔的多示例学习的演示。**左边：**正包中包含正示例（方形和菱形）和负示例（圆形）。本章中，假设正示例分布于不同的子类别，划分于不同的椭圆当中。本章方法的目的是学习一个最大化间隔的分类器去将正包中的属于不同子类别以及负类别的示例分开。**右边：**负包中包含的全部都是负示例。

本章将在MIL的框架内，提出一种全新的码本学习的方法。首先我们可以很自然地将图像和MIL的表示融合在一起，将一幅图作为一个包，图像中的一个图像块（或一个区域）当做一个示例。对于不同类别的图像，将其中感兴趣的一类当做正样本，其余的作为负样本。对每一幅图像，如果它被标为正样本，那么图像中的至少有一个图像块被认为是正样本；相反，如果它被标为了负类图像，图像中的所有块都应当被当做负样本。以Scene 15数据集[19]中的图像为例，如果将高速公路这一类的图像作为正样本，那么山峰、乡村等类别的图像就成为了负样本；值得注意的是，属于天空的图像块在正负样本中都出现了，那么它们将被认为是负的示例。如图5-2所示，假设正的图像块来自于不同的子类别，负的图像块来自于一个负的类别。本章的目的在于学习一个最大化间隔的分类器将这些正的子类别以及负

类别之间互相分开，并展示这些学习得到的分类器可以用于图像的分类。对于传统的码本学习中的称呼，这些分类器在本章中被称为是通用码字（Generalized code，G-code）。这个的码本学习问题涉及两个子问题：（1）已知示例的子类别，学习区分型的混合模型，（2）已知区分型混合模型，决定示例自动分类（每个图像块应属于哪一子类）。这样看来，MIL可以很直接的解决上述问题。因此，接下来，本章将先给出一个原始的解决方法，然后给出本章提出的最大化间隔的多示例学习问题的详细的定义和推导。

5.2.2 一个原始的解决方案

基于上述思路，我们需要将多示例学习和最大化间隔分类器，如，SVM，结合起来。回顾多示例学习的一个经典解决方法——由Andrews等人提出的mi-SVM算法[51]。在这个方法中，利用示例标记来学习SVM和利用SVM来推测示例标记这两个步骤迭代的进行，最终mi-SVM可以从正包中选择出来正示例。在选择出来正示例之后，可以采用传统的k-means码本学习方法来学习一个码本。具体的算法流程在算法5-1中展示，这个解决方案是一个原始的解决方案。这个解决方案中多示例学习和混合模型学习是完全分离的，二者互不影响。这显然不是一个最优的解决方案，因为事实上多示例学习和混合模型学习是两个相辅相成的步骤。在接下来的内容中将尝试把多示例学习和混合模型学习两个步骤结合在一个统一的框架下，提出可用于码本学习的最大化间隔的多示例学习。

算法 5-1 一个原始的多示例码本学习的解决方案。

给定正包和负包，完成下面两个步骤：

MIL步骤: 在输入的正包和负包上运行mi-SVM算法，将正示例和负示例分开，并保留学习得到的正示例。

聚类步骤: 将上一步得到的正示例作为输入，运行k-means算法，得到最终的码本。

5.2.3 最大化间隔的多示例学习的数学推导

最大化间隔的多示例学习力求将多示例学习和码本学习融合在一起，并直接最大化不同子类别之间的间隔。为了实现这个目的，本章采用多类的SVM[189]并不使用任何复杂的核函数。不失一般性，简单的线性分类器表示为$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$。每一个子类对应于一个特定的线性分类器。由于多类SVM的灵活性，在学习阶段可以很容易的让不同子类的分类器相互竞争。在本章中，子类标记被视作为隐变量，对于每一个示例的子标记表示为$z_{ij} \in \{0, 1, \ldots, K\}$。$z_{ij} = k \in \{1, \ldots, K\}$表示示例$\mathbf{x}_{ij}$属于

正类中的第k个子类。反之，$z_{ij}=0$表示x_{ij}属于负类。另外，还需要定义一个权重矩阵，$\mathbf{W}=[\mathbf{w}_0,\mathbf{w}_1,\ldots,\mathbf{w}_K],\mathbf{w}_k\in\mathbf{R}^{d\times 1},k\in\{0,1,\ldots,K\}$，来表示$K+1$个线性分类器，其中每一列表示一个线性分类器，具体来讲$\mathbf{w}_k,k>0$表示第k子类的线性分类器模型，$\mathbf{w}_0$表示负类的线性分类器模型。这样示例$\mathbf{x}_{ij}$的子类标记由下式表示：

$$z_{ij}=\arg\max_k \mathbf{w}_k^T\mathbf{x}_{ij} \tag{5-1}$$

根据以上定义，接下来给出最大化间隔的多示例学习的目标方程：

$$\begin{aligned}\min_{\mathbf{W},z_{ij}}\quad & \sum_{k=0}^{K}\parallel \mathbf{w}_k\parallel^2+\lambda\sum_{ij}\max(0,1+\mathbf{w}_{r_{ij}}^T\mathbf{x}_{ij}-\mathbf{w}_{z_{ij}}^T\mathbf{x}_{ij})\\ \text{s.t.}\quad & \text{if } Y_i=1,\sum_j z_{ij}>0,\text{and if } Y_i=0,z_{ij}=0,\end{aligned} \tag{5-2}$$

其中$r_{ij}=\arg\max_{k\in\{0,\ldots,K\},k\neq z_{ij}}\mathbf{w}_k^T x_{ij}$为一个辅助变量。在公式(5-2)中，第一项$\sum_{k=0}^{K}\parallel \mathbf{w}_k\parallel^2$是边界的规范化；第二项是多类的铰链损失（hinge loss），表示为$\ell(\mathbf{W};(\mathbf{x}_{ij},z_{ij}))$。

$$\ell(\mathbf{W};(\mathbf{x}_{ij},z_{ij}))=\sum_{ij}\max(0,1+\mathbf{w}_{r_{ij}}^T\mathbf{x}_{ij}-\mathbf{w}_{z_{ij}}^T\mathbf{x}_{ij}) \tag{5-3}$$

其中，参数λ用来控制以上两项之间的权重。损失函数$\ell(\mathbf{W};(\mathbf{x}_{ij},z_{ij}))$ 实习了直接最大化$K+1$个子类别的边界。公式(5-2)中的约束条件等同于多示例学习中的约束条件：$\sum_j z_{ij}>0\Leftrightarrow\sum_j y_{ij}>0$ 并且$z_{ij}=0\Leftrightarrow y_{ij}=0$。

以上最大化间隔的多示例学习问题引出了一个非凸的优化问题。但这个问题是一个半凸（semi-convex）问题[23]，因为一旦隐变量（正包中的示例的子标记）的值确定，这个问题就会变成一个凸优化问题。在文献[23]中，一个“坐标下降（oordinate descend）”的方法，能够处理类似问题。但是相比较而言，公式(5-2) 中的问题的更加困难，因为在这个目标方程中，正示例的数目未知。

5.3 最大化间隔的多示例学习的优化

本节将给出公式(5-2)中的优化问题给出一个解决方案，完成最大化间隔的多示例学习。首先，可以将训练集表示为$\mathcal{D}=\{X_1,\ldots,X_n\}$ 其中包含用于训练的所有正包和负包。然后对每一个示例定义一个权重系数，定义示例权重（instance weight）如下：

$$\begin{aligned}p_{ij}&=\text{sigmoid}(\max_{k\in\{1,\ldots,K\}}\left(\mathbf{w}_k^T\mathbf{x}_{ij}-\mathbf{w}_0^T\mathbf{x}_{ij}\right)/\sigma)\\&=\left(1+\exp\left(-\max_{k\in\{1,\ldots,K\}}\left(\mathbf{w}_k^T\mathbf{x}_{ij}-\mathbf{w}_0^T\mathbf{x}_{ij}\right)/\sigma\right)\right)^{-1}\end{aligned} \tag{5-4}$$

p_{ij}表示的是样本的“正度（positiveness）”。它是由最大的那个正的子类与负类的SVM决策值之差决定的，具体写为$\max_{k\in\{1,\ldots,K\}}\left(\mathbf{w}_k^T x_{ij}-\mathbf{w}_0^T x_{ij}\right)$。其中Sigmoid函数用来将SVM决策值之差映射到$(0,1)$之间。σ是一个用作归一化的参数。

算法 5-2 最大化间隔的多示例学习的优化策略。

输入值: 正包，负包以及子类数目K。

初始化: 对于负包中的示例，我们设置将其子标记设为0，$z_{ij}=0$。对于正包中的示例，我们采用k-means算法将所有的示例划分到K个子类别中。所有示例的权重设置为1，$p_{ij}=1$。

对于以下两个步骤，迭代执行N次（在实验中，通常N设置为5）：

优化W: 根据示例的权重p_{ij}，我们从所有的正包中的示例中采样p^s的样本，并采用所有的负包中的负示例，形成一个训练集$\mathcal{D}'$。训练集$\mathcal{D}'$中的所有样本的子标记都是已知的，因此我们可以训练一个多类SVM并得到码本$\mathbf{W}$,

$$\min_{\mathbf{W}} \quad \sum_{k=0}^{K} \parallel \mathbf{w}_k \parallel^2 +\lambda \sum_{ij} \max(0, 1+\mathbf{w}_{r_{ij}}^T \mathbf{x}_{ij} - \mathbf{w}_{z_{ij}}^T \mathbf{x}_{ij})$$

其中$\mathbf{x}_{ij}\in\mathcal{D}'$ and $r_{ij}=\arg\max_{k\in\{0,\ldots,K\},k\neq z_{ij}}\mathbf{w}_k^T\mathbf{x}_{ij}$。

更新p_{ij}和z_{ij}: 在$\mathbf{W}$已知的情况下，我们更新示例权重p_{ij} 和隐变量z_{ij}。具体做法是对正包中的所有示例重复以下两个步骤:

1) 根据公式(5-4)来更新示例权重p_{ij}。
2) 按照如下公式更新隐变量:

$$z_{ij}=\arg\max_{k\in\{1,\ldots,K\}}\left(\mathbf{w}_k^T x_{ij}-\mathbf{w}_0^T x_{ij}\right)$$

输出值: 学习得到的码本$\mathbf{W}$。

在下一步中，将采用一种随机化的坐标下降算法来解决(5-2) 中的问题，这个算法摘要如算法5-2。首先建立一个新的训练集$\mathcal{D}'$，$\mathcal{D}'$是通过从原始的训练集$\mathcal{D}$中根据样本权重p_{ij}采样得到的。因为隐变量只对于正包中的示例起作用，负包中的样本的子标记已确定——均为负类，所以将负包中的所有示例都放在$\mathcal{D}'$中。另外，对于每一个正包，不全部使用其中的示例，仅采用p^s比率的样本。在最初，因为没有任何先验信息，每个包中所有示例的权重都是一样的。在完成示例采样步骤之后，训练集$\mathcal{D}_0'$ 将被用作训练一个标准的多类SVM分类器f_0。这样就完成了算法5-2中的算法的**优化W**步骤。一旦得到了f_0，便可以在原始的正包中采用f_0来进行**更新p_{ij}和z_{ij}**的步骤。接下来的一轮中，从正包中根据跟新之后的示例权重采样p^s 比率的正示例来形成新一轮的训练集$\mathcal{D}_1'$并训练得到f_1。如此两个步骤不断迭代，直至规定的迭代次

数已经到达。

在算法5-2中的算法中，根据示例的positiveness 确保了正包中的一部分示例具有正的子标记——这满足了公式(5-2)中的多示例约束。另外，通过采样步骤也可以显著的提高整个优化算法的效率。

5.4 基于最大化间隔的多示例码本的图像表示

学习得到的码本由一系列线性分类器（G-code）组成。这些线性分类器将不同的子类别分开。类似于Li Fei-Fei等人提出的物体组（object bank）[184] 方法中的思路，我们可以利用学习得到的G-code分类器来进行图像的表示。如图5-3所示，假设在整个图像数据集中存在M个类别的图像，码本学习方法是对于这M个类别单独进行的。在针对其中的某一类训练的时候，当前类别中的图像被认为是正样本（正包），其它的类别的所有图像被认为是负样本（负包）。利用本章中的最大化间隔的多示例学习，可以学习得到$K+1$个G-code分类器。

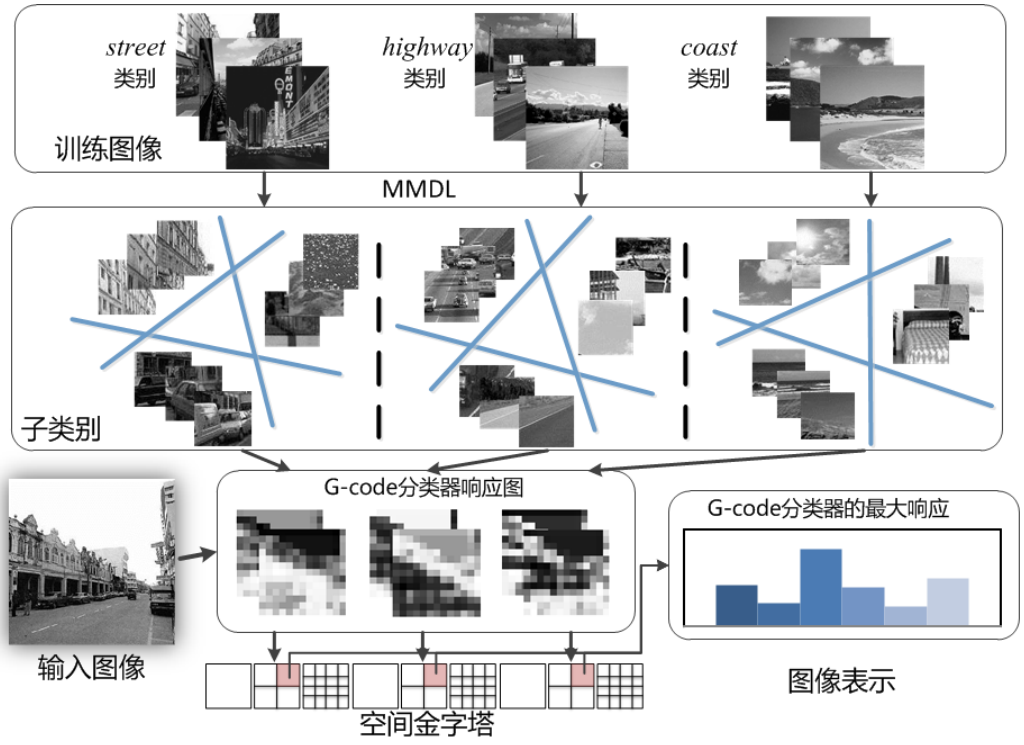


图 5-3 基于最大化间隔的多示例码本的图像表示的演示。给定一个图像集合（如第一行所示），其中包含属于多个类别的图像。基于最大化间隔的多示例学习方法，将这些图像中的图像块聚类到不同的子类，同时学习得到G-code分类器，即码本。对于每一个学习得到的子类（包含正子类和负类），在第二张中显示了一些例子。对于一幅输入图像，利用G-code分类器的响应图和空间金字塔来建立图像表示（在第三行显示）。

对于给定的一幅输入图像，从图像中密集的提取图像块作为局部特征。假设$\mathbf{x}$是一个局部特征向量，第k个G-code在$\mathbf{x}$上的响应表示为$\mathbf{w}_k^T\mathbf{x}, k \in \{0, 1, \ldots, K\}$。这样，对于输入图像我们可以得到每一个局部特征对于第k个G-code的响应，从而形成一个响应图（response map）。对于每一个响应图，一个三层的空间金字塔[19]被用来建模图像中局部特征间的空间关系。这样三层的空间金字塔将图像分割为$(1^2 + 2^2 + 4^2) = 21$个空间栅格。在每一个空间栅格中，对于每一个G-code分类器，在响应图上取最大的响应作为这个区域的特征。这样，在每一个空间栅格中，就可以得到一个$M \times (K + 1)$维的特征。最终的图像表示是将所有栅格中的特征连接在一起得到一个$21 \times M \times (K + 1)$的特征。

可以发现，采用G-code进行图像表示的复杂度很低。对于每一个特征的编码仅仅需要完成一个点乘操作。对比其它的一些标准的图像分类系统[190]中的编码方法，本章的算法的复杂度显然十分的低。另外，由于MMDL方法的码本远远小于传统的码本，这进一步提高了本章图像表示方法的效率。对于高层的视觉应用，如图像分类，在众多标准测试集上本章的图像表示方法能够取得十分优异的图像分类结果。

5.5 实验

在实验中，本章中所提出的最大化间隔的多示例码本学习方法简称为MMDL。下面将以此介绍数据集、实验设置、并分别按照不同数据集给出图像识别的结果和分析。

数据集 实验中选择了四个广泛使用的数据集上测试本章中所提出的MMDL方法在图像分类上的性能，包括场景图像（15 Scene数据集[19], MIT 67 Indoor数据集[191] ），行为图像（UIUC Sports数据集[192]）和物体图像（Caltech 101数据集[89]）。这四个数据集的详细信息和测试准则如下：

- 15 Scene数据集：它包含4485图像，这些图像被分到了15个类别，每个类别中存在200到400幅图像，平均的图像大小为300×250。依照这个数据集上的标准测试方法[19]，对于每个类别，实验中随机采用100幅图像进行训练，剩余的图像用作测试。
- MIT 67 Indoor数据集：这个数据集中包含来自于67个不同类共计15620幅室内图像。对于这个数据集，有固定的训练和测试数据划分，对于每一类，大概采用80幅图像作为训练，其余图像作为测试。
- UIUC Sports数据集：这个数据集中被分为8个运动行为类别，如划船、打羽毛球、攀岩等。依照这个数据集上的标准测试方法[192]，对于每个类别，实验中随机采用60幅图像进行训练，剩余的图像用作测试。
- Caltech 101数据集：这个数据集中有9144幅图像，来自于101 物体类别和一个背景类别。每个类别中的图像数量在31到800不等。根据这个数据集上的通用

设定，实验中从每一类中随机的选择30幅图像用作训练，剩余的图像用作测试。

对于15 Scene，UIUC Sports和Caltech 101数据集，随机运行实验5次，并计算5次运行的平均精度和标准差。MIT 67 Indoor数据集上的训练样本和测试样本固定，因此不用随机划分训练测试样本，直接运行一次就可以得到实验结果。

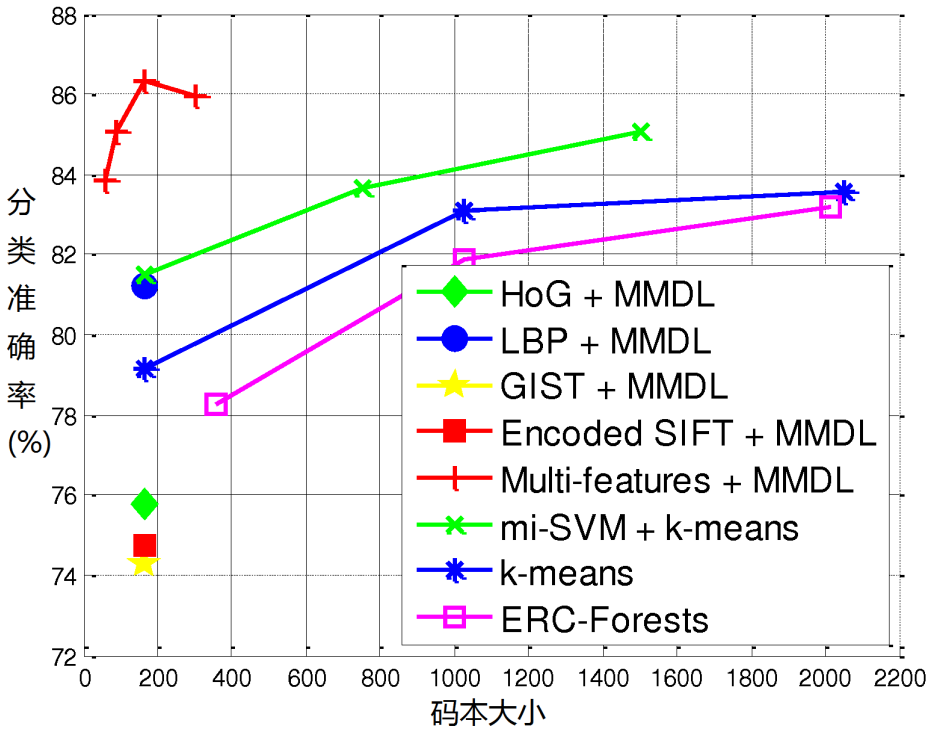


图 5-4 15 Scene dataset数据集上的图像分类准确率（%），横坐标表示码本的大小，纵坐标表示不同方法的性能。

实验设置 对于每一幅图像，在x，y方向上，每间隔16个像素，实验中提取48×48, 72×72, 和96×96三个尺度上的图像块。对于每一个图像块，将其缩放到48×48，然后计算不同特征来描述这个图像块。所采用的特征包含HOG，LBP，GIST[193]，编码的SIFT和LAB颜色直方图。对于HOG和LBP特征，采用开源计算机视觉算法库VLFeat[194]中的实现，它们的维度分别是279维和522维。对于GIST特征，采用其提出者的实现[193]，GIST特征的维度是256。对于编码的SIFT，每间隔6个像素计算大小为16×16的图像块的SIFT特征，再采用k-means算法将其量化到100个不同的区间，每个SIFT 被分配到最近的一个区间，这样就可以得到一个100维的直方图作为特征。对于LAB颜色直方图，在三个颜色通道中各计算16维的直方图，最终形成一个48维的特征向量。这5个不同的特征单独归一化后会拼接成为一个1205维的向量，

最终经过ℓ_2归一化成为最终的局部特征的表示。在MMDL中，权重系数λ设置为1。算法5-2中的优化算法中的迭代次数设置为5，采样率p^s设置为0.7，归一化系数σ设置为0.5，在步骤“优化W”中，采用开源SVM库LibLinear[79]来解决这个多类SVM问题。每个数据集中的训练图像都用来学习码本。根据章节5.4中的算法，对每张图像建立图像表示。最终，LibLinear再次被采用进行最终的图像分类。

表 5-1 15 Scene dataset数据集上的图像分类准确率（%），以及不同方法所采用的码本的大小。

方法	分类准确率(%)	码本大小
Object Bank[184]	80.90	2400
Lazebnik等人的方法[19]	81.10±0.30	200
Zhu等人的方法[180]	82.4±0.7	1024
Yang等人的方法[13]	80.40±0.45	1024
Kernel Descriptors[195]	**86.70±0.40**	1000
MMDL	86.35±0.45	**165**

5.5.1 自然场景的图像分类

在这个实验中，采用15 Scene数据集来进行自然场景的分类，将本章中提出的MMDL算法与k-means码本学习，极度随机化的聚类森林（extremely randomized clustering forests，ERC-Forests）[174]，章节5.2.2中的原始的优化方法,以及其它的一些图像分类方法进行比较。

在图5-4中，X轴表示k-means码本大小，或者G-code的数目；Y轴表示图像分类的正确率（%）。HOG，LBP，GIST和编码的SIFT（encoded SIFT）分别采用MMDL方法编码，在采用MMDL编码时，每类采用11个G-code，总共165个G-code。由于这个数据集上只提供了灰度图像，因此LAB颜色直方图特征在这个数据集上未被采用。实验结果是：LBP的正确率（81.23%）远高于HOG（75.7%），编码的SIFT（74.74%）和GIST（74.27%）的正确率。将这四种特征融合，分类准确率达到了显著的提升，为86.35%。

在采用全部的四种特征情况下，另外测试了传统的k-means码本学习方法，ERC-Forests码本学习方法，以及基线方法，即章节5.2.2中的mi-SVM+k-means方法。采用k-means，ERC-Forests和mi-SVM+k-means方法学习得到的码本都采用局部约束线性编码（LLC）[14]来进行编码，LLC方法是目前最流行且最有效的编码方法，在很多性能优异的系统中都得到了使用。然后采用文献[14]中的框架来完成图像分类。在图5-4中，可以观察到ERC-Forests方法弱于k-means方法，基线方法（mi-SVM

+ k-means）比直接采用k-means方法要好，因为基线方法采用多示例学习挖掘了图像中具有区分性的特征。尽管如此，它的性能要逊于MMDL方法。mi-SVM + k-means方法采用1500个码本，可以得到85.06%的平均正确率，然而MMDL方法仅采用165个G-code就能达到86.35%的平均正确率。

在表5-1中，将MMDL方法与之前的一些图像分类方法进行了比较。值得注意的是，Lazebnik等人的方法[19]和Bo等人的Kernel Descriptors方法[195]采用了非线性SVM进行图像分类。Object Bank方法[184]和Yang等人的方法[13]同本章的方法一样，采用线性分类器来进行图像分类。可以观察到，本章所提出的MMDL方法的性能十分接近于业界最优的性能，但是采用了非常小的码本和简单快速的线性SVM分类器来进行图像分类。

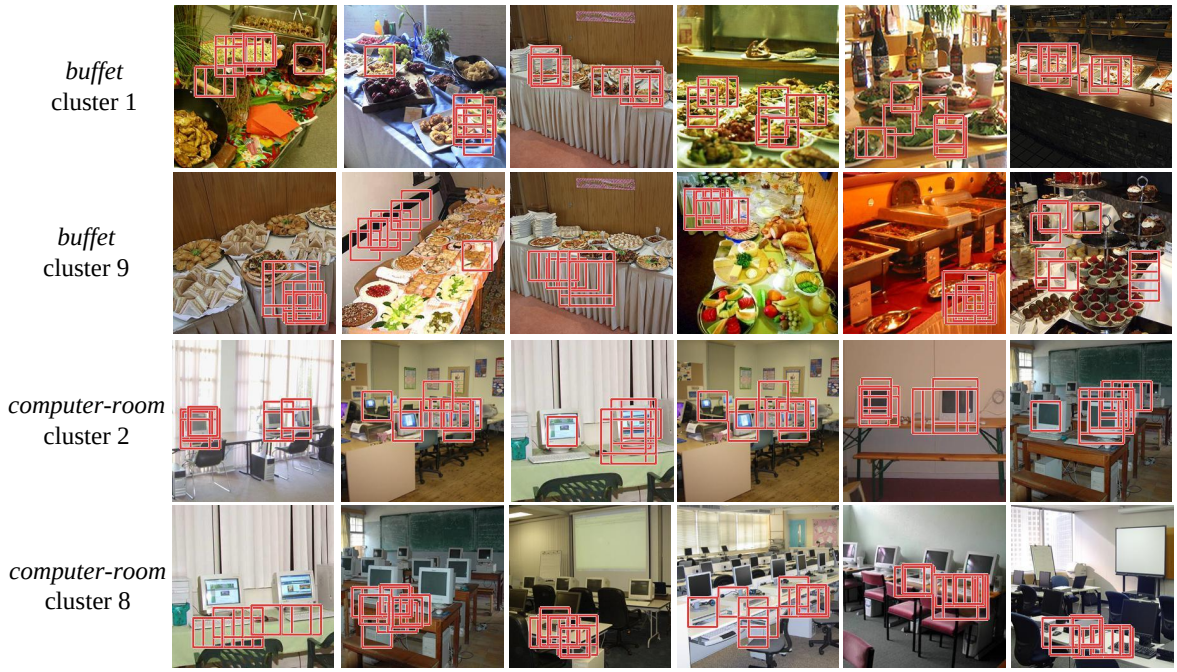


图 5-5 采用MMDL方法在不同类别上学习得到的一些有意义的子类（上图中记为cluster）。每一行展示了一个子类，红色的方框标注了那些G-code分类器所对应的SVM方程值大于0的图像位置。

5.5.2 室内场景图像分类

这个实验将在在MIT 67 Indoor数据集上测试MMDL在室内场景图像的分类上的性能。对于67类中的每一类，学习11个G-code，其中10个对应于正子类，1个对应于负类，总共有737个G-code分类器。图5-5中显示了MMDL算法学习到的一些比较有意义的子类别。在上面两行中显示了buffet这个类别中学习到的两个子类，其中第1个子类对应于beffet中的餐盘，第9个子类对应于桌布；在下面两行中显示

了computer-room这个类别中学习到的两个子类，其中第2个子类对应于电脑的显示器，第8个子类对应于电脑桌。这些很有语义信息的子类别是在只给定图像级标记的情况下，由MMDL算法推断出来的，这充分地说明了MMDL的有效性。

表 5-2 MIT 67 Indoor数据集上图像分类性能的比较

方法	分类准确率(%)
ROI+Gist[191]	26.5
MM-scene[196]	28.0
Centrist[197]	36.9
Object Bank[184]	37.6
DPM[198]	30.4
RBoW[178]	37.93
Disc. Patches[179]	38.1
SPMSM[199]	44.0
LPR[200]	44.84
MMDL	**50.15**

表5-2中将MMDL方法在这个数据集上的平均分类准确率与之前发表的一些方法进行了比较。相对于之前的传统的场景分类方法[191,196,197]，MMDL方法的性能明显更为优异。因此我们可以更注重于同三个采用中层图像表示（id-level image representations）的方法进行比较，它们分别是DPM方法[198]，RBoW方法[178] 和Discriminative Patches方法[179]。DPM，RBoW，Discriminative Patches和本章提出的MMDL方法均采用了图像类别的标注信息。在Discriminative Patches方法中，作者将学习得到的Discriminative Patches与DPM，Gist-color，SP 结合到一起得到了49.4%的准确率。MMDL方法取得了更高的准确率，另外，如果结合其它的一些方法，MMDL可以取得更好的结果。

5.5.3 体育行为图像分类

这个实验将在UIUC Sports进行进行体育行为的识别，对于这个数据集中8个体育活动类别，分别采用各个类别的一部分图像进行训练，其余的图像用作测试。对于每一个类别，仍然采用11个G-code，这样总共用88个G-code分类器来作为图像表示。如表5-3所示，尽管采用了很小的码本，但是MMDL的结果明显优于Object Bank方法（需要额外的物体标注）和两个最新的方法LPR[200]和SPMSM[199]。另外，还对比了Wu等人提出的一种基于直方图交叉核的码本学习方法[201]。

表 5-3 UIUC Sports数据集上的图像分类正确率

方法	分类准确率(%)
Li等人的方法[192]	73.4
Wu等人的方法[201]	84.3
Object Bank[184]	76.3
SPMSM[199]	83.0
LPR[200]	86.25
MMDL	**88.47±2.32**

5.5.4 物体图像分类

除了上述的场景图像分类和体育行为图像分类，我们还将在一个标准的图像测试集，Caltech 101数据集，上进行了物体图像分类。在这个实验中，对于每一个物体类别，学习6个G-code分类器，其中5个对应于正的子类，1个对应于负类，这样总共有612个G-code 分类器用来进行图像的表示。

表 5-4 Caltech 101数据上的图像分类准确率

方法	分类准确率(%)
NBNN[202]	73.0
LLC[14]	73.4±0.5
CDBN[203]	65.5
Kernel Descriptors[195]	76.±0.7
LP-β[92]	77.7±0.3
TS-MKL[204]	0.772
MMDL	**78.18±0.17**

在表5-4中，首先将本章提出的MMDL方法同两个基于多尺度稠密SIFT的方法进行比较，其中一个采用简单的贝叶斯最近邻方法（Naive Bayesian Nearest Neighbor，NBNN）[202]，另外一个为LLC方法[14]。MMDL方法的平均精度大概比它们高5个百分点，原因在MMDL方法可以学习到有意义的中层特征，并且融合多种图像特征。然后，表5-4中还比较了两个特征学习方法，分别是卷积深度置信网络（Convolutional Deep Belief Networks，CDBN）[203]和核描述子[195]。最后，比较了两个基于多核学习（Multiple Kernel Learning，MKL）的特征融合算法，包括经典的LP-β方法[92]和最近的TS-MKL方法[204]，在这两个方法中，39个不同核被应用于5个不同的特征上。MMDL方法只是简单地对4个不同的特征学习一个权重系数，

获得了比LP-β和TS-MKL略好的性能。

5.6 本章小结

本章中引入了一种的多示例学习方法：传统的多示例学习当中均只有正示例和负示例两种，本章中根据真实图像数据中正示例的实际分布，将正示例划分成不同的子类别中，提出了最大化间隔的多示例学习。并且，本章首次将多示例学习引入到图像码本的学习当中。大量的图像数据上的实验充分地展示了本章方法的有效性，同其它的一些图像识别方法相比，本章中提出的图像码本尺寸小，建立图像表达速度快的优点。

6 总结与展望

6.1 全文总结

本文就物体识别这个计算机视觉领域的核心问题，结合计算机视觉的原理和机器学习理论，对于图像分类、物体检测和物体发现等任务提出了四个行之有效的方法，即轮廓片段包方法（BoCF）、扇形形状方法（FSM）、基于低秩优化的子空间学习方法和最大化间隔的多示例学习方法（MMDL）。其中BoCF方法和MMDL方法在对应的形状识别和图像分类两个应用上不仅仅在精度上达到了业界先进水平，而且在速度上也领先于之前的同类方法，能够直接应用于工业应用中。例如，本文提出的MMDL方法已经应用于智能交通系统中的车辆的识别软件当中，现已有成熟的产品，并在申请相关专利。除了提出了一些能够解决实际问题的方法，本文在理论上对于计算机视觉和机器学习的研究也做出了贡献。

（1）基于部件的形状表示。本文第2章和第3章中的方法均采用了基于部件的形状表示。在第2章BoCF方法中，采用由DCE算法得到的多尺度轮廓片段作为形状的部件，利用形状码本和特征的编码来完成不同形状之间的相似部件匹配。然后采用SVM来选择有区分性的形状部件来有效的完成形状的分类。同理，第5章的方法MMDL也可以和BoCF结合，用来选择有区分性的形状部件。BoCF方法的突出贡献在于它对于形状给出了一个简洁有效的向量形式的表示。在第3章FSM方法中，本文着重于建立一个结构化的基于部件的形状模型，对比BoCF，FSM更加显性的描述了形状不同部件的空间关系，从而实现更加精确的物体检测，相对于DPM等方法只能定位物体的包围盒，FSM可以更精确的来定位物体的轮廓。同DPM等方法的模型学习设定一样，FSM方法能够在弱监督情况下学习得到基于部件的物体模型，其模型学习思路是通过匹配来获得物体部件之间的对应关系。对比DPM方法中的隐变量SVM方法，本文的思路更加快捷，不依赖于负样本，且泛化能力更强。BoCF方法和FSM方法都证实了采用基于部件的形状表示是一个行之有效的物体识别思路。

（2）多示例学习。本文第4章和第5章的研究内容同多示例学习紧密相关，为这个机器学习领域的热点问题做出了一系列贡献。传统的多示例学习中，示例的标记只有正负两类。本文的研究中根据实际的应用情况突破的了这个两类假设。第一种情况出现在第4章的物体发现任务中，只有正包的存在，而没有负包可以利用，因此本文提出单类的多示例学习。为解决这个单类多示例学习问题，本文提出了基于低秩优化的子空间学习方法，假设正示例存在于一个低维的子空间中，利用低秩优化排除掉离群点，即正包中的负示例。该方法得到的目标方程是一个凸优化问题，比之前的多示例学习的形式化更加优美简洁，最后采用ADMM方法快速的求得全局最优解，在物体发现任务上取得了优异的实验结果。第二种情况出现在第5章的码本学

习任务中，针对图像中物体的不同部件的不同外观，第5章的MMDL方法将正示例划分到不同的子类以对应物体不同部件的不同外观，这种多示例学习方法可以称为多类的多示例学习。MMDL方法针对这种情况提出了一个基于最大化间隔的目标方程，并提出随机化的坐标下降方法进行求解。MMDL方法学习到的图像码本富含高层语义，简洁有效，在图像分类上取得了优异的性能。MMDL的码本学习思路对于弱监督情况下图像中语义挖掘很有借鉴意义，例如美国UIUC大学的著名的Thomas Huang教授研究组就利用了MMDL方法在众包数据中挖掘有效的图像特征来完成精确的车辆分类[205]。

本文中的研究方法有效的解决了物体识别中的关键问题，且富有启发性。同时，本人将实验代码发布在网上同其他的研究人员分享。本文的FSM方法和MMDL方法分别在发表两年和一年之内均获得了10次的Google Scholar引用，取得了本领域的研究人员的广泛关注。

6.2 研究展望

如绪论中图1-5所描述，本文围绕物体识别中的形状建模和弱监督学习展开了诸多研究工作，除了实线表示的已经在本文中得到研究的工作，还包括三个虚线表示的未完成的工作，有待下一步探索。（1）基于形状模型的物体发现，利用形状特征的有效性和物体的自身结构特点，尝试自动的从图像中发掘出存在形变的共同物体。（2）基于弱监督学习的物体检测，利用弱监督学习图像中层特征，然后结合滑动窗口来完成物体检测。（3）采用最大化间隔的多示例学习来学习基于部件的物体模型，在给定物体包围盒的情况下，利用MMDL学习得到有区分性的物体部件，并在学习过程中约束物体部件之间的空间一致性，最后得到类似于DPM模型的物体检测器。

物体识别是计算机视觉中一个任重道远的问题。在未来的研究中，沿着本文研究的探索思路，结合当前的研究现状，未来以下方面值得深入的研究。

（1）弱监督情况下的深度学习。目前的深度学习算法一般采用两种方式，要么基于强监督学习，如CNN，要么基于无监督学习，如SAE（Stacked Autoencoder）和RBM（Restricted Boltzmann Machine），没有在弱监督情况下进行深度学习的研究工作。鉴于弱监督学习擅长于在样本标记较弱的情况下挖掘图像的语义，以及深度学习强大的表达能力，将弱监督学习同深度学习结合可以利用更多的训练样本。如在ILSVRC比赛中，如果可以将图像分类的弱标注数据用于物体检测，这样就可以使得物体检测获得多于之前近百倍的训练样本。

（2）部件级别（Part Level）的深度学习。深度学习算法并没有一个有效的机制去处理物体识别中由于形变带来的困难。然而基于部件的物体模型则擅长于处理物体的形变。将物体部件的位置当做隐变量放置于深度学习当中，采用EM算法的思路去推测物体部件的位置。这个过程可以看作是部件级别的深度学习，用来学习对于

物体形变具有更强处理能力的深度学习模型，这是一个值得探索的方向。

（3）结合上下文（Context）的物体识别模型。虽然采用上下文来进行物体识别在本文中没有得到深入的探讨，但它在物体识别中有着至关重要的作用，能够显著提升物体识别的性能。采用本文中的研究方法如MMDL来表示上下文特征，并将其同物体模型结合在一起进行物体识别是一个值得研究的问题。

致　谢

这是一段纯粹的、疯狂的、值得怀念的岁月。在华科九年多的时间里，我历经了一个从崇尚技术的本科生到攀登学术高峰的博士生的转变，深深的沉醉在学术研究的快乐中，或熬夜改论文，或做梦想idea，我都乐在其中。读博士的过程不但让我掌握了如何创新性地去解决尖端的科研问题，同时也历练了我的性格，让我更加坚韧和睿智，为我今后的工作打好了基础。在此，我要向所有在读博期间给予我关怀和帮助的老师、同学、朋友和亲人表示感激。

感谢我的导师刘文予教授！刘老师是我的良师益友，有刘老师作为我的导师是我这二十年学生生涯中的最大幸运！难以想象一位导师能够支持自己的学生在博士阶段在校外进行三年多的访问学习，感谢刘老师的支持和信任！感谢刘老师多年来的悉心指导和辛勤栽培！刘老师治学严谨，厚德载物，是学生人生道路上的榜样，在今后的工作中，学生定当追求更多的成绩以回报恩师！

感谢白翔教授！白老师是将我带入研究殿堂的引路人，带领我完成了本科毕业设计，投稿ICCV 2009并被录用。我们一起去异国他乡开会，一起熬夜科研，一起喝酒聚餐，整整五年的工作和生活构建了我们深厚的友谊。白老师的智慧和拼搏精神让我敬佩，愿在今后的工作和生活中，我们能够精诚合作，一起取得更大的成就。

感谢加州大学圣地亚哥分校的屠卓文教授！能够在博士生涯一开始同屠老师这样一位在领域内如此德高望重、成就斐然的著名学者合作是我巨大的荣幸。整个博士阶段，屠老师对于我的指导也是无微不至的，在Temple University交流的时候，屠老师就和John Wright一起打越洋电话同我讨论采用稀疏表示的来进行物体检测，之后带我在微软亚洲研究院视觉计算组实习，接下来还推荐我到UCLA Alan Group访问研究。屠老师在研究方面的敏锐嗅觉，对于研究的透彻理解让我敬佩万分，待我亦师亦友，愿今后的研究生涯中能够继续跟随屠老师的脚步，做出更多更有影响力的工作。

感谢Temple University的Longin Latecki教授！在费城访问的这段时间是我研究成果出的最快的阶段。Latecki教授对于研究的极大热情深深感染了我，我们在Wachman Hall无数次的讨论让我们共同构想的ideas一个一个变成文章。通过这段时间的历练，我从对于研究懵懵懂懂的状态到开始对于研究有了自己的理解。近些年，Latecki教授虽然饱受疾病的困扰，但我们直接关于研究的讨论却从未间断，祝愿您早日康复！

感谢加州大学洛杉矶大学Alan Yuille教授！Yuille教授让我有机会去接触到美国顶级的研究学府，感受其中的浓厚学术氛围。作为我们领域的泰山北斗式的学者，Yuille教授却是那样的平易近人，总会孜孜不倦的给我讲解各种有趣的知识，让我敬

仰万分。

感谢上海科技大学马毅教授！马老师的高风亮节，学术造诣，天分和勤奋都让我十分的惊叹。非常荣幸能够同马老师一起合作开展研究，一同探讨问题，这些经历将使我终身受用。

感谢Dian团队刘玉教授！刘老师的是我本科的导师，在我读博期间也给了非常多的帮助，是一位十分可爱可敬的老师。Dian团队期间在技术和性格两个方面的历练对于我之后的学术研究都有巨大的帮助。

感谢刘海容师兄、杨兴炜师兄、李劝男师兄、姚聪、沈为、周瑜、马天阳、马佳义、郭晓杰、陈攀、邱卫超！我们之间不仅仅学术上的合作者，更是交心的好兄弟！

感谢媒体与通信实验室的冯镔老师、田臣老师、蒋洪波老师、王邦老师、江南老师在学术和工作上对于我的关心和指导！

感谢微软亚洲研究院的王井东研究员，王宝元研究员，MIT的张峥东和哥伦比亚大学的孙举同学，感谢你们在学术交流中给予我的指导和帮助！

感谢媒体与通信实验室的陈珺、刘俊涛、王波、邓天生、何洋、姚志均、张拯、柏松、石葆光、章成全、朱盈盈、饶聪、王跃铭、王长涛、蔡超、叶小琼、朱山、易思华、陈娇艳、周全、王军伟、何乐、危俊、张新浩、鲁勤、申辰、贺芳姿、熊祎、段雄、朱卓暾、戴一桥等同学，感谢你们在我整个研究生生涯的帮助，和你们一起工作和玩耍非常开心。

感谢在Temple University的汪成倞、李淑莎、王允生、易萌、杜亮、喻秀文、舒乐、成二康、吴毅、郎海涛、兰亮、王庄、楼强等朋友，费城的14个月是我博士阶段最为轻松快乐的时光，费城一年四季的美丽风光我将永远铭记。

感谢在微软亚洲研究院的杨杨、苗丹、刘衡、殷大伟、常时雨、朱俊、吴佳俊、王历伟等朋友，有幸和你们在微软亚洲研究院这么一个福地相逢，非常开心可以和你们共同奋斗。

感谢在UCLA的刘小白师兄、Xiaochen、王建宇、陈先捷、夏方婷、任洲、Liang-Chieh Chen，Boyan Bonev，George Papandreou，Roozbeh Mottagh等朋友，感谢你们在UCLA对于我的帮助。

感谢微软亚洲研究院授予我“微软学者”奖学金的激励和资助。

感谢我的家人！感谢关心我的人和我关心的人！

参考文献

[1] Lowe D G. Distinctive Image Features From Scale-invariant Keypoints. Int'l J. of Comp. Vis., 2004, 60(2):91–110.

[2] Mikolajczyk K, Schmid C. Scale & affine invariant interest point detectors. International journal of computer vision, 2004, 60(1):63–86.

[3] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005, 27(10):1615–1630.

[4] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. in: Proceedings of Computer Vision–ECCV 2006, pages 404–417. Springer, 2006.

[5] Rublee E, Rabaud V, Konolige K, et al. ORB: an efficient alternative to SIFT or SURF. in: Proceedings of Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, 2564–2571.

[6] Yang X, Cheng K. Local Difference Binary for Ultra-fast and Distinctive Feature Description. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 1(1):1–10.

[7] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4):509–522.

[8] Sivic J, Zisserman A. Robust text-indep. speaker ident. using Gaussian mixture speaker models. in: Proceedings of Video Google: A Text Retrieval Approach to Object Matching in Videos, 2003, 1470-1477.

[9] Csurka G, Dance C, Fan L, et al. Visual categorization with bags of keypoints. Workshop on Statistical Learning in Computer Vision, ECCV, 2004, 2004.

[10] Nister D, Stewenius H. Scalable recognition with a vocabulary tree. in: Proceedings of Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. IEEE, 2006, 2161–2168.

[11] Wang X, Wang B, Bai X, et al. Max-margin multiple-instance dictionary learning. in: Proceedings of Proceedings of The 30th International Conference on Machine Learning, 2013, 846–854.

[12] Jiang Z, Zhang G, Davis L. Submodular dictionary learning for sparse coding. in: Proceedings of CVPR. IEEE, 2012, 3418–3425.

[13] Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. IEEE, 2009, 1794–1801.

[14] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. IEEE, 2010, 3360–3367.

[15] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification. in: Proceedings of Europe Conference on Computer Vision, 2010. Springer, 2010, 143–156.

[16] Wang X, Bai X, Liu W, et al. Feature context for image classification and object detection. in: Proceedings of Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, 961–968.

[17] Arandjelovic R, Zisserman A. All about VLAD. in: Proceedings of Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, 1578–1585.

[18] Grauman K, Darrell T. The pyramid match kernel: Discriminative classification with sets of image features. in: Proceedings of International Conference on Computer Vision, 2005. IEEE, 2005, 1458–1465.

[19] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006. IEEE, 2006, 2169–2178.

[20] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. in: Proceedings of Proc. of CVPR, 2005, 886-893.

[21] Viola P A, Jones M J. Robust Real-Time Face Detection. Int'l J. of Comp. Vis., 2004, 57(2):137–154.

[22] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. J. of Comp. and Sys. Sci., 1997, 55(1):119–139.

[23] Felzenszwalb P, Girshick R, McAllester D, et al. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9):1627–1645.

[24] Zhang J, Huang K, Yu Y, et al. Boosted local structured hog-lbp for object localization. in: Proceedings of Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, 1393–1400.

[25] Dean T, Ruzon M A, Segal M, et al. Fast, accurate detection of 100,000 object classes on a single machine. in: Proceedings of Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, 1814–1821.

[26] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. in: Proceedings of Advances in Neural Information Processing Systems, 2012, 2012, 1106–1114.

[27] Deng J, Dong W, Socher R, et al. ImageNet: A Large-Scale Hierarchical Image Database. in: Proceedings of CVPR09, 2009.

[28] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. Neural computation, 1989, 1(4):541–551.

[29] Wang T, Wu D J, Coates A, et al. End-to-end text recognition with convolutional neural networks. in: Proceedings of Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012, 3304–3308.

[30] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on, 2012, 20(1):30–42.

[31] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. arXiv preprint arXiv:1405.0312, 2014.

[32] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524, 2013.

[33] Marr D. Vision. W.H. Freeman and Co. San Francisco, 1982.

[34] Ling H, Jacobs D. Shape classification using the inner-distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2):286–299.

[35] Bai X, Wang X, Latecki L J, et al. Active Skeleton for Non-rigid Object Detection. ICCV, 2009.

[36] Felzenszwalb P, Schwartz J. Hierarchical matching of deformable shapes. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007. IEEE, 2007, 1–8.

[37] Thayananthan A, Stenger B, Torr P H, et al. Shape context and chamfer matching in cluttered scenes. in: Proceedings of Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. IEEE, 2003, I–127.

[38] Liu M Y, Tuzel O, Veeraraghavan A, et al. Fast directional chamfer matching. in: Proceedings of CVPR, 2010, 1696–1703.

[39] Ferrari V, Fevrier L, Jurie F, et al. Groups of adjacent contour segments for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(1):36–51.

[40] Shotton J, Blake A, Cipolla R. Multi-Scale Categorical Object Recognition Using Contour Fragments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 90(7):1270–1281.

[41] Opelt A, Pinz A, Zisserman A. A boundary-fragment model for object detection. ECCV, 2006.

[42] Yarlagadda P, Ommer B. From meaningful contours to discriminative object shape. in: Proceedings of ECCV, pages 766–779. Springer, 2012.

[43] Ma T, Latecki L. From partial shape matching through local deformation to robust global shape similarity for object detection. in: Proceedings of CVPR, june, 2011, 1441 -1448.

[44] Riemenschneider H, Donoser M, Bischof H. Using partial edge contour matches for efficient object category localization. ECCV, 2010, pages 29–42.

[45] Fischler M, Elschlager R. The representation and matching of pictorial structures. IEEE Tran. on Computers, 1973, C-22:67–92.

[46] Felzenszwalb P F, Huttenlocher D P. Pictorial Structures for Object Recognition. IJCV, 2005, 61(1):55–79.

[47] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. in: Proceedings of Proc. of 10th Int'l Conf. on Machine Learning, San Francisco, 2001, 282-289.

[48] Vapnik V. Estimation of dependences based on empirical data. Springer-Verlag, 1982.

[49] Viola P A, Platt J, Zhang C. Multiple instance boosting for obj. detect. in: Proceedings of Proc. of NIPS, 2005.

[50] Babenko B, Yang M H, Belongie S. Visual tracking with online multiple instance learning. in: Proceedings of Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, 983–990.

[51] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. in: Proceedings of Advances in Neural Information Processing Systems. MIT Press, 2003, 561–568.

[52] Breiman L. Random Forests. Mach. Learn., 2004, 45:5–32.

[53] Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes. in: Proceedings of CVPR, 2009, 1778–1785.

[54] Kim W, Kim Y. A region-based shape descriptor using Zernike moments. Signal Processing: Image Communication, 2000, 16(1):95–102.

[55] Zhang D, Lu G. Generic Fourier descriptor for shape-based image retrieval. in: Proceedings of IEEE International Conference on Multimedia and Expo. IEEE, 2002, 425–428.

[56] Mokhtarian F, Abbasi S, Kittler J, et al. Efficient and robust retrieval by shape content through curvature scale space. Series on Software Engineering and Knowledge Engineering, 1997, 8:51–58.

[57] Adamek T, O'Connor N. A multiscale representation method for nonrigid shapes with a single closed contour. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(5):742–753.

[58] Alajlan N, El Rube I, Kamel M, et al. Shape retrieval using triangle-area representation and dynamic space warping. Pattern Recognition, 2007, 40(7):1911–1920.

[59] Xu C, Liu J, Tang X. 2D shape matching by contour flexibility. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(1):180–186.

[60] Sun K, Super B. Classification of contour shapes using class segment sets. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005. IEEE, 2005, 727–733.

[61] Bai X, Liu W, Tu Z. Integrating contour and skeleton for shape classification. in: Proceedings of International Conference on Computer Vision Workshops (ICCV Workshops), 2009. IEEE, 2009, 360–367.

[62] Daliri M, Torre V. Robust symbolic representation for shape recognition and retrieval. Pattern Recognition, 2008, 41(5):1782–1798.

[63] Daliri M, Torre V. Shape recognition based on kernel-edit distance. Computer Vision and Image Understanding, 2010, 114:1097–1103.

[64] Wang B, Shen W, Liu W, et al. Shape classification using tree-unions. in: Proceedings of International Conference on Pattern Recognition, 2010. IEEE, 2010, 983–986.

[65] Torsello A, Hancock E. Learning shape-classes using a mixture of tree-unions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(6):954–967.

[66] Erdem A, Tari S. A similarity-based approach for shape classification using Aslan skeletons. Pattern Recognition Letters, 2010, 31(13):2024–2032.

[67] Siddiqi K, Shokoufandeh A, Dickinson S, et al. Shock graphs and shape matching. International Journal of Computer Vision, 1999, 35(1):13–32.

[68] Torsello A, Hancock E. A skeletal measure of 2D shape similarity. Computer Vision and Image Understanding, 2004, 95(1):1–29.

[69] Bai X, Latecki L. Path similarity skeleton graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(7):1282–1292.

[70] Baseski E, Erdem A, Tari S. Dissimilarity between two skeletal trees in a context. Pattern Recognition, 2009, 42(3):370–385.

[71] Eslami S, Heess N, Winn J. The shape boltzmann machine: a strong model of object shape. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012. IEEE, 2012, 406–413.

[72] Latecki L, Lakämper R. Convexity rule for shape decomposition based on discrete contour evolution. Computer Vision and Image Understanding, 1999, 73(3):441–454.

[73] Moosmann F, Nowak E, Jurie F. Randomized clustering forests for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(9):1632–1646.

[74] Yang J, Yu K, Huang T. Supervised translation-invariant sparse coding. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010. IEEE, 2010, 3517–3524.

[75] Duda R, Hart P, Stork D. Pattern Classification and Scene Analysis. 1995.

[76] Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500):2323–2326.

[77] Serre T, Wolf L, Poggio T. Object recognition with features inspired by visual cortex. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005. IEEE, 2005, 994–1000.

[78] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. The Journal of Machine Learning Research, 2002, 2:265–292.

[79] Fan R, Chang K, Hsieh C, et al. LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research, 2008, 9:1871–1874.

[80] Latecki L, Lakamper R, Eckhardt T. Shape descriptors for non-rigid shapes with a single closed contour. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2000. IEEE, 2000, 424–429.

[81] Leibe B, Schiele B. Analyzing appearance and contour based methods for object categorization. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2003. IEEE, 2003, II–409.

[82] Daliri M, Torre V. Shape recognition and retrieval using string of symbols. in: Proceedings of International Conference on Machine Learning and Applications, 2006. IEEE, 2006, 101–108.

[83] Attalla E, Siy P. Robust shape similarity retrieval based on contour segmentation polygonal multiresolution and elastic matching. Pattern Recognition, 2005, 38(12):2229–2241.

[84] Lim K L, Galoogahi H. Shape classification using local and global features. in: Proceedings of Pacific-Rim Symposium on Image and Video Technology, 2010, 2010, 115–119.

[85] Söderkvist O. Computer vision classification of leaves from swedish trees: [PhD Dissertation]. Linköping, 2001.

[86] Hu R, Jia W, Ling H, et al. Multiscale Distance Matrix for Fast Plant Leaf Recognition. IEEE transactions on image processing, 2012.

[87] Hu R, Jia W, Zhao Y, et al. Perceptually motivated morphological strategies for shape retrieval. Pattern Recognition, 2012.

[88] Wang J, Bai X, You X, et al. Shape matching and classification using height functions. Pattern Recognition Letters, 2012, 33(2):134–143.

[89] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding, 2007, 106(1):59–70.

[90] Arbelaez P, Maire M, Fowlkes C, et al. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5):898–916.

[91] Kovesi P D. MATLAB and Octave Functions for Computer Vision and Image Processing. Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia, 2013. Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[92] Gehler P, Nowozin S. On feature combination for multiclass object classification. in: Proceedings of International Conference on Computer Vision, 2009. IEEE, 2009, 221–228.

[93] Zhang H, Berg A C, Maire M, et al. SVM-KNN: Discriminative nearest neighbour classification for visual category recognition. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006. IEEE, 2006, 2126–2136.

[94] Zhang Y, Jiang Z, Davis L S. Learning Structured Low-rank Representations for Image Classification. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, 2013, 676 - 683.

[95] Shaban A, Rabiee H, Farajtabar M, et al. From Local Similarity to Global Coding; An Application to Image Classification. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, 2013, 2794 - 2801.

[96] Canny J. A computational approach to edge detection. IEEE Trans. on PAMI, 1986, 8(6).

[97] Martin D, Fowlkes C, Malik J. Learning to Detect Natural Image Boundaries Using Local Brightness, Color and Texture Cues. IEEE Trans. on Pattern Analysis and Machine Learning, 2004, 26(5):530–549.

[98] Dollár P, Tu Z, Belongie S. Supervised Learning of Edges and Object Boundaries. in: Proceedings of Proc. of CVPR, New York, June, 2006.

[99] Dollár P, Zitnick C L. Structured forests for fast edge detection. in: Proceedings of Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013, 1841–1848.

[100] Xiaofeng R, Bo L. Discriminatively trained sparse code gradients for contour detection. in: Proceedings of Advances in neural information processing systems, 2012, 584–592.

[101] Ferrari V, Tuytelaars T, Gool L V. Object Detection by Contour Segment Networks. ECCV, 2006.

[102] Toshev A, Taskar B, Daniilidis K. Shape-Based Object Detection via Boundary Structure Segmentation. International Journal of Computer Vision, 2011, 99(5814):123–146.

[103] Yang X, Liu H, Latecki L J. Contour-based object detection as dominant set computation. Pattern Recognition, 2012, 45(5):1927–1936.

[104] Everingham M, Van Gool L, Williams C K I, et al. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.

[105] Kovesi P D. Matlab and octave functions for computer vision and image processing. 2008.

[106] Crandall D J, Huttenlocher D P. Weakly Supervised Learning of Part-Based Spatial Models for Visual Object Recognition. in: Proceedings of ECCV, 2006, 16-29.

[107] Kokkinos I, Yuille A. Inference and learning with hierarchical shape models. International Journal of Computer Vision, 2011, 93(2):201–225.

[108] Crandall D, Felzenszwalb P, Huttenlocher D. Spatial priors for part-based recognition using statistical models. in: Proceedings of CVPR, 2005, 10–17.

[109] Zhang Z. Iterative point matching for registration of free-form curves and surfaces. International journal of computer vision, 1994, 13(2):119–152.

[110] Rath T M, Manmatha R. Word image matching using dynamic time warping. in: Proceedings of Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. IEEE, 2003, II–521.

[111] Cormen T H, Leiserson C E, Rivest R L, et al. Introduction to Algorithms. MIT Press, 2nd edition, 2001.

[112] Zhang Z, Ganesh A, Liang X, et al. TILT: transform invariant low-rank textures. International Journal of Computer Vision, 2012, 99(1):1–24.

[113] Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Arxiv preprint arXiv:1009.5055, 2010.

[114] Gumbel E J, Greenwood J A, Durand D. The circular normal distribution: Theory and tables. Journal of the American Statistical Association, 1953.

[115] Bai X, Yang X, Latecki L, et al. Learning context-sensitive shape similarity by graph transduction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(5):861–874.

[116] Ferrari V, Jurie F, Schmid C. From images to shape models for object detection. International Journal of Computer Vision, 2010.

[117] Srinivasan P, Zhu Q, Shi J. Many-to-one Contour Matching for Describing and Discriminating Object Shape. CVPR, 2010.

[118] Frey B, Dueck D. Clustering by passing messages between data points. Science, 2007, 315(5814):972–976.

[119] Maji S, Malik J. A max-margin hough tranform for object detection. CVPR, 2009.

[120] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. CVPR, 2008.

[121] Lu C, Adluru N, Ling H, et al. Contour based object detection using part bundles. Computer Vision and Image Understanding, 2010, 114(7):827–834.

[122] Kontschieder P, Riemenschneider H, Donoser M, et al. Discriminative Learning of Contour Fragments for Object Detection. in: Proceedings of BMVC, 2011, 4–1.

[123] Zhu Q, Wang L, Wu Y, et al. Contour Context Selection for Object Detection: A Set-to-Set Contour Matching Approach. European Conference on Computer Vision, 2008.

[124] Jenatton R, Obozinski G, Bach F. Structured sparse principal component analysis. in: Proceedings of International Conference on Artificial Intelligence and Statistics, 2010, 366-373.

[125] Wright J, Yang A, Ganesh A, et al. Robust Face Recognition via Sparse Representation. IEEE Transactions Pattern Analysis and Machine Intelligence, 2009, 31(2):210–227.

[126] Wagner A, Wright J, Ganesh A, et al. Towards a Practical Face Recognition System: Robust Registration and Illumination via Sparse Representation. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, 597-604.

[127] Elhamifar E, Vidal R. Sparse subspace clustering. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, 2790-2797.

[128] Liu G, Lin Z, Yu Y. Robust subspace segmentation by low-rank representation. in: Proceedings of Proceedings of the 26th International Conference on Machine Learning, 2010, 663-670.

[129] Luo D, Nie F, Ding C, et al. Multi-subspace representation and discovery. Machine Learning and Knowledge Discovery in Databases, 2011, pages 405–420.

[130] Favaro P, Vidal R, Ravichandran A. A closed form solution to robust subspace estimation and clustering. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, 1801-1807.

[131] Zhu J, Wu J, Wei Y, et al. Unsupervised Object Class Discovery via Saliency-Guided Multiple Class Learning. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012, 3218-3225.

[132] Gabay D, Mercier B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers & Mathematics with Applications, 1976, 2(1):17–40.

[133] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning, 2011, 3(1):1–122.

[134] Dietterich T, Lathrop R, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 1997, 89(1-2):31–71.

[135] Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Royal Statist. Soc. Series B, 1977, 39:1–8.

[136] Yu C, Joachims T. Learning structural svms with latent variables. in: Proceedings of Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, 1169–1176.

[137] Wang X, Zhang Z, Ma Y, et al. One-Class Multiple Instance Learning via Robust PCA for Common Object Discovery. in: Proceedings of Asian Conference on Computer Vision, 2012, 246-258.

[138] Lerman G, McCoy M B, Tropp J A, et al. Robust computation of linear models, or How to find a needle in a haystack. CoRR, 2012, abs/1202.4044.

[139] Russell B, Freeman W, Efros A, et al. Using multiple segmentations to discover objects and their extent in image collections. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006, 1605-1614.

[140] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation. J. of Machine Learning Res., 2003, 3:993–1022.

[141] Grauman K, Darrell T. Unsupervised learning of categories from sets of partially matching image features. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006, 19-25.

[142] Lee Y, Grauman K. Foreground focus: Unsupervised learning from partially matching images. International Journal of Computer Vision, 2009, 85(2):143–166.

[143] Candes E, Li X, Ma Y, et al. Robust Principal Component Analysis? Journal of the ACM, 2011, 58(3):1–37.

[144] Horn R A, Johnson C R. Matrix Analysis. Cambridge University Press, 2012.

[145] Tao M, Yuan X. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM Journal on Optimization, 2011, 21(1):57–81.

[146] Shiqian Ma L X, Zou H. Alternating Direction Methods for Latent Variable Gaussian Graphical Model Selection. Neural Computation, 2013, 25(8):2172–2198.

[147] Georghiades A, Belhumeur P, Kriegman D. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions Pattern Analysis and Machine Intelligence, 2001, 23(6):643–660.

[148] Everingham M, Van Gool L, Williams C K I, et al. The PASCAL Visual Object Classes Challenge (VOC) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index .html, 2011.

[149] Feng J, Wei Y, Tao L, et al. Salient object detection by composition. in: Proceedings of International Conference on Computer Vision, 2011, 1028–1035.

[150] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object Detection with Discriminatively Trained Part Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9):1627–1645.

[151] Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12):2037–2041.

[152] Everingham M, Zisserman A, Williams C K I, et al. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006.

[153] Everingham M, Van Gool L, Williams C K I, et al. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[154] Jain V, Learned-Miller E. FDDB: A Benchmark for Face Detection in Unconstrained Settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[155] Ferrari V, Tuytelaars T, Van Gool L. Object detection by contour segment networks. in: Proceedings of European Conference on Computer Vision, 2006, 14-28.

[156] Deselaers T, Alexe B, Ferrari V. Weakly Supervised Localization and Learning with Generic Knowledge. International Journal of Computer Vision, 2012, 100(3):275–293.

[157] Pandey M, Lazebnik S. Scene recognition and weakly supervised object localization with deformable part-based models. in: Proceedings of IEEE International Conference on Computer Vision. IEEE, 2011, 1307–1314.

[158] Chum O, Zisserman A. An exemplar model for learning object classes. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007, 1–8.

[159] Lampert C, Blaschko M, Hofmann T. Efficient subwindow search: a branch and bound framework for object localization. IEEE Transactions Pattern Analysis and Machine Intelligence, 2009, pages 2129–2142.

[160] Rahmani R, Goldman S A, Zhang H, et al. Localized content based image retrieval. in: Proceedings of ACM SIGMM international workshop on Multimedia information retrieval. ACM, 2005, 227–236.

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

[162] Dietterich T, Lathrop R, Lozano-Perez T. Solving the multiple-instance problem with axis parallel rectangles. Artificial Intelligence, 1997, 89:31–71.

[163] Chen Y, Bi J, Wang J Z. MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12):1931–1947.

[164] Zhang Q, Goldman S A. EM-DD: An Improved Multiple-Instance Learning Technique. in: Proceedings of Advances in Neural Information Processing Systems. MIT Press, 2001, 1073–1080.

[165] Wang H Y, Yang Q, Zha H. Adaptive p-posterior mixture-model kernels for multiple instance learning. in: Proceedings of Proceedings of the 25th Annual International Conference on Machine learning. ACM, 2008, 1136–1143.

[166] Zhou Z H, Sun Y Y, Li Y F. Multi-instance learning by treating instances as non-iid samples. in: Proceedings of Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, 1249–1256.

[167] Deselaers T, Ferrari V. A Conditional Random Field for Multiple-Instance Learning. in: Proceedings of Proceedings of the 26th International Conference on Machine Learning, 2010, 287-294.

[168] LeCun Y, Huang F, Bottou L. Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting. in: Proceedings of Proc. of CVPR, June, 2004.

[169] Wright J, Yang A, Ganesh A, et al. Robust Face Recognition via Sparse Representation. IEEE Trans. PAMI, 2009, 31(2).

[170] Hinton G E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 2006, 18(7):1527–1554.

[171] Duda R, Hart P, Stork D. Pattern Classification and Scene Analysis. John Wiley and Sons, 2000.

[172] Jurie F, Triggs B. Creating Efficient Codebooks for Visual Recognition. in: Proceedings of ICCV, 2005, 604-610.

[173] Lazebnik S, Raginsky M. Supervised learning of quantizer codebooks by information loss minimization. IEEE Tran. PAMI, 2009, 31:1294–1309.

[174] Moosmann F, Nowak E, Jurie F. Randomized Clustering Forests for Image Classification. IEEE Trans. Pattern Anal. Mach. Intell., 2008, 30(9):1632–1646.

[175] Yang L, Jin R, Sukthankar R, et al. Unifying discriminative visual codebook generation with classifier training for object category recognition. in: Proceedings of Proc. of CVPR, 2008.

[176] Mairal J, Bach F, Ponce J. Task-driven dictionary learning. arXiv preprint arXiv:1009.5358, 2010.

[177] Winn J, Criminisi A, Minka T. Object categorization by learned universal visual dictionary. in: Proceedings of ICCV, 2005, 1800–1807.

[178] Parizi S, Oberlin J, Felzenszwalb P. Reconfigurable models for scene recognition. in: Proceedings of CVPR. IEEE, 2012, 2775–2782.

[179] Singh S, Gupta A, Efros A A. Unsupervised Discovery of Mid-Level Discriminative Patches. in: Proceedings of ECCV, 2012.

[180] Zhu J, Zou W, Yang X, et al. Image Classification by Hierarchical Spatial Pooling with Partial Least Squares Analysis. in: Proceedings of British Machine Vision Conference, 2012.

[181] Pechyony D, Vapnik V. On the theory of learning with privileged information. in: Proceedings of NIPS, 2010.

[182] Parikh D, Grauman K. Relative attributes. in: Proceedings of ICCV, 2011, 503–510.

[183] Bourdev L, Malik J. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. in: Proceedings of ICCV, 2009.

[184] Li L, Su H, Xing E, et al. Object bank: A high-level image representation for scene classification and semantic feature sparsification. NIPS, 2010, 24.

[185] Dollár P, Babenko B, Belongie S, et al. Multiple component learning for object detection. ECCV, 2008, pages 211–224.

[186] Xu Y, Zhu J, Chang E, et al. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. in: Proceedings of CVPR, 2012, 964–971.

[187] Zhang M L, Zhou Z H. M3MIML: A maximum margin method for multi-instance multi-label learning. in: Proceedings of ICDM, 2008, 688–697.

[188] Zhang D, Wang F, Si L, et al. M3IC: maximum margin multiple instance clustering. in: Proceedings of IJCAI, 2009, 1339–1344.

[189] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. The Journal of Machine Learning Research, 2002, 2:265–292.

[190] Chatfield K, Lempitsky V, Vedaldi A, et al. The devil is in the details: an evaluation of recent feature encoding methods. in: Proceedings of BMVC, 2011.

[191] Quattoni A, A.Torralba. Recognizing Indoor Scenes. in: Proceedings of CVPR, 2009.

[192] Li L, Fei-Fei L. What, where and who? classifying events by scene and object recognition. in: Proceedings of ICCV, 2007.

[193] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 2001, 42(3):145–175.

[194] Vedaldi A, Fulkerson B. VLFeat: An Open and Portable Library of Computer Vision Algorithms, 2008.

[195] Bo L, Ren X, Fox D. Kernel descriptors for visual recognition. NIPS, 2010, 7.

[196] Zhu J, Li L, Fei-Fei L, et al. Large margin learning of upstream scene understanding models. NIPS, 2010, 24.

[197] Wu J, Rehg J. CENTRIST: A visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8):1489–1501.

[198] Pandey M, Lazebnik S. Scene recognition and weakly supervised object localization with deformable part-based models. in: Proceedings of ICCV. IEEE, 2011, 1307–1314.

[199] Kwitt R, Vasconcelos N, Rasiwasia N. Scene Recognition on the Semantic Manifold. in: Proceedings of ECCV, 2012.

[200] Sadeghi F, Tappen M. Latent Pyramidal Regions for Recognizing Scenes. in: Proceedings of ECCV, 2012.

[201] Wu J, Rehg J. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. in: Proceedings of ICCV, 2009, 630–637.

[202] Boiman O, Shechtman E, Irani M. In defense of nearest-neighbor based image classification. in: Proceedings of CVPR. IEEE, 2008, 1–8.

[203] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. in: Proceedings of ICML. ACM, 2009, 609–616.

[204] Kumar A, Niculescu-Mizil A, Kavukcuoglu K, et al. A Binary Classification Framework for Two-Stage Multiple Kernel Learning. in: Proceedings of ICML, 2012, 1295–1302.

[205] Xu N, Wang J, Wang Z, et al. An ontological bagging approach for image classification of crowd-sourced data. in: Proceedings of Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on. IEEE, 2014, 1–5.

附录 1　攻读学位期间发表的学术论文

[1] Xinggang Wang, Bin Feng, Xiang Bai, Wenyu Liu, Login Latecki. Bag of Contour Fragments for Robust Shape Classification. Pattern Recognition, Volume 47, Issue 6, June 2014, Pages 2116-2125.（SCI，5年影响因子：3.153，Google Scholar引用次数为4）

[2] Xinggang Wang, Zhengdong Zhang, Yi Ma, Xiang Bai, Wenyu Liu, Zhuowen Tu. Robust Subspace Discovery via Relaxed Rank Minimization. Neural Computation, Vol. 26, No. 3，April 2014, Pages 611-635.（SCI，5年影响因子：2.121，Google Scholar引用次数为1）

[3] Xinggang Wang, Baoyuan Wang, Xiang Bai, Wenyu Liu, Zhuowen Tu. Max-Margin Multiple Instance Dictionary Learning. International Conference on Machine Learning (ICML), 2013, Atlanta, USA.（CCF A类会议，Google Scholar引用次数为12）

[4] Xinggang Wang, Xiang Bai, Tianyang Ma, Wenyu Liu, Longin Latecki. Fan Shape Model for Object Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2012, Providence, USA. （CCF A类会议，Google Scholar引用次数为11)

[5] Xinggang Wang, Xiang Bai, Xingwei Yang, Wenyu Liu, Longin Jan Latecki. Maximal Cliques that Satisfy Hard Constraints with Application to Deformable Object Model Learning. Neural Information Processing Systems Conference (NIPS), 2011, Granada, Spain. （CCF B类会议，Google Scholar引用次数为6）

[6] Xinggang Wang, Xiang Bai, Wenyu Liu, Longin Latecki. Feature Context for Object Detection and Image Classification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2011, Colorado Spring, USA. (CCF A类会议，Google Scholar引用次数为37)

[7] Xinggang Wang, Zhengdong Zhang, Yi Ma, Xiang Bai, Wenyu Liu, Zhuowen Tu. One-Class Multiple Instance Learning via Robust PCA for Common Object Discovery. Asian Conference on Computer Vision (ACCV), 2012, Daejeon, Korea. （CCF C类会议，Google Scholar引用次数为1）

[8] Xiaojie Guo, Xinggang Wang, Liang Yang, Xiaochun Cao, Yi Ma. Robust Foreground Detection Using Smoothness and Arbitrariness Constraints. European Conference on Computer Vision (ECCV), Zurich, September, 2014

[9] Xiang Bai, Xinggang Wang, Longin Jan Latecki, Wenyu Liu, Zhuowen Tu. Active Skeleton for Non-rigid Object Detection. IEEE International Conference on Computer

Vision (ICCV), 2009, Kyoto, Japan.

[10] Weichao Qiu, Xinggang Wang, Xiang Bai, Alan Yuille, Zhuowen Tu. Scale-space SIFT Flow, IEEE Winter Conference on Applications of Computer Vision (WACV), 2014, Steamboat Springs CO, USA.

[11] Song Bai, Xinggang Wang, Cong Yao, Xiang Bai. Multiple Stage Residual Model for Accurate Image Classification. Asian Conference on Computer Vision (ACCV), 2014 (Accepted).

[12] Wei Shen, Xinggang Wang, Cong Yao, Xiang Bai. Shape Recognition by Combining Contour and Skeleton into a Mid-level Representation. China Conference on Pattern Recognition (CCPR), Shanghai, Chian, 2014.

[13] Yingying Zhu, Xinggang Wang, Cong Yao, Xiang Bai. Traffic sign classification using two-layer image representation. International Conference on Image Processing (ICIP), 2013, Melbourne, Astralia.

[14] Yueming Wang, Xinggang Wang, Shaojun Zhu, Xiang Bai, Wenyu Liu. Adjacent Coding for Image Classification. International Conference on Pattern Recognition (ICPR), 2012, Tsukuba Science City, Japan.

[15] Quannan Li, Xinggang Wang, Wei Wang, Yuan Jiang, Zhi-Hua Zhou, Zhuowen Tu. Disagreement-Based Multi-System Tracking. ACCV Workshop of Detection and Tracking in Challenging Environments, 2012.

[16] Chen Duan, Xinggang Wang, Shuiming Shu, Changwei Jing, Huawei Chang. Thermodynamic design of Stirling engine using multi-objective particle swarm optimization algorithm. Energy Conversion and Management, Volume 84, August 2014, Pages 88–96. （SCI，5年影响因子：3.604）

[17] Xiang Bai, Cong Rao, Xinggang Wang*. Shape Vocabulary: A Robust and Efficient Shape Representation for Shape Matching. IEEE Transactions on Image Processing (TIP). Accepted. （*通讯作者）（SCI，5年影响因子：3.925）

[18] Xiang Bai, Bo Wang, Xinggang Wang, Wenyu Liu, Zhuowen Tu. Co-Transduction for Shape Retrieval. European Conference on Computer Vision (ECCV), 2010, Crete, Greece.

[19] Bo Wang, Xiang Bai, Xinggang Wang, Wenyu Liu, Zhuowen Tu. Object Recognition using Junctions. European Conference on Computer Vision (ECCV), 2010, Crete, Greece.

[20] Meng Yi, Tatyana Nuzhnaya, Vasileios Megalooikonomou, Xinggang Wang, Longin Jan Latecki, Mark Kohn, Robert Steiner. Lung Image Classification using Locality-Constrained Linear Coding. STMI 2012: MICCAI Workshop on Sparsity Techniques in Medical Imaging.

附录 2　攻读学位期间申请专利列表

1. “基于成本敏感的自适应增强的人脸认证方法”，发明人：彭勇，白翔，王兴刚，沈为，王波。专利申请号：201010183540 X。
2. “一种运用最大子图的基于局部模型的物体检测方法”，发明人：白翔，王兴刚，申辰，刘文予。专利申请号：201210248431.0。
3. “一种基于匹配的车辆颜色识别方法和系统”，发明人：陈瑞军，白翔，陈攀，王兴刚，肖可伟。专利申请号：201210248431.0。

附录 3　攻读博士学位期间获得的奖励

1. 2012年”微软学者奖”（全亚洲仅10 人获奖）
2. 2014年华中科技大学优秀博士论文基金
3. 2012年华中科技大学博士国家奖学金
4. 2013年华中科技大学三好研究生奖
5. NIPS 2012, Travel Grant
6. IEEE ICCV 2009, Travel Grant

附录 4　攻读博士学位期间的学术服务

以下期刊和会议的审稿人:

- IEEE Transactions on Cybernetics
- Pattern Recognition
- IEEE Intelligent Transportation Systems Transactions and Magazine
- Computer Vision and Image Understanding
- ICMD 2011
- CVPR 2014
- ECCV 2014