

谨以此论文献给我的导师和亲人！

————— 王如晨

基于多核学习的浮游生物图像分类研究

学位论文答辩日期: _____

指导教师签字: _____

答辩委员会成员签字: _____

独 创 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含未获得 _____ (注：如没有其他需要特别声明的，本栏可空) 或其他教育机构的学位或证书使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 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，并同意以下事项：

1. 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磁盘，允许论文被查阅和借阅。
2. 学校可以将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、汇编学位论文。同时授权清华大学“中国学术期刊(光盘版)电子杂志社”用于出版和编入CNKI《中国知识资源总库》，授权中国科学技术信息研究所将本学位论文收录到《中国学位论文全文数据库》。

(保密的学位论文在解密后适用本授权书)

学位论文作者签名： 导师签字：
签字日期： 年 月 日 签字日期： 年 月 日

基于多核学习的浮游生物图像分类研究

摘要

浮游生物是海洋中生物的基本能量来源，构成了海洋食物链的基础，因此其丰富程度会影响海洋生态系统的平衡。并且，浮游生物对环境的改变较为敏感，专家们也可以利用其这一特点研究环境的变化。此外，大量有害浮游生物繁殖还会造成污染。因此，监测海洋浮游生物的分布和丰富度对海洋生态系统研究以及海洋环境保护等工作具有重要意义。

早期对浮游生物的监测是通过人工采集浮游生物样本，并由专业人士借助显微镜等设备进行分类统计实现。这个过程不仅耗时耗力，而且专业要求高，因此导致浮游生物监测效率低下。为了解决这一问题，人们先后研制了浮游生物图像采集系统和自动识别系统，这两个系统可以高效的采集并识别大量浮游生物图像，他们的出现不仅提高浮游生物监测的效率，也降低了成本和专业需求。然而，现有浮游生物图像分类系统的泛化能力和分类性能有待提高。因此，本文以提高浮游生物图像分类系统整体性能为目标，研究并提出了基于多核学习的浮游生物图像分类系统，以提高浮游生物分类的准确率和适用范围，使其更好的应用于浮游生物监测中，主要工作如下：

- 构建不同的浮游生物数据集。为了降低所设计图像分类系统的数据集偏见，提高整个系统的泛化能力，我们搜集了不同机构和设备采集的浮游生物图像，用其构建不同的数据集进行实验，这些数据集中既包含浮游动物数据集，也包含浮游植物数据集。
- 研究从多角度对浮游生物的形态特征进行分析，提出结合多种有效特征的浮游生物描述方法。首先，根据浮游生物的形态，选用适合的统计方法提取几何灰度等特征；此外，将目标检测与识别领域的经典特征提取方法应用于浮游生物描述，例如局部二值模式、梯度方向直方图、内距离形状上下文等算法；然后，针对提取特征中的冗余部分，采用特征选择算法将其去除，为每个数据集保留最优的特征组合。
- 提出基于多核学习的浮游生物图像分类方法。对于提取的不同种类的浮游生物特征，分别为其设计适合的核函数，利用多核学习算法进行融合，从而充分发挥每种特征在分类中的积极作用。
- 设计对比实验，构建合理的评价体系，对提出的基于多核学习的浮游生物图像分类系统的性能进行评价。首先，根据目前应用较好的浮游生物图像分类方法设计基准实验，作为评价分类器性能的基准；然后在基准实验的基础上设计特征对比实验，对特征提取部分的性能进行评价；最后，采用本文设计

的分类系统进行实验，与之前实验结果进行对比评价。

实验结果表明，本文提出的基于多核学习的浮游生物图像分类系统具有较好的分类性能和泛化能力，在不同浮游生物数据集上都取得了较好的分类结果。

关键词： 浮游生物，多核学习，特征提取，核函数，特征融合

Plankton image classification based on multiple kernel learning

Abstract

Plankton is the main source of food for organisms in the ocean and forms the base of marine food chain. So the abundance of it will influence the ocean ecological balance. In addition, plankton is very sensitive to environment changes, thus it can be used to study the changing environment. And harmful plankton bloom pollutes the marine environment. Therefore, the study of plankton abundance and distribution is important, in order to understand environment change and protect marine ecosystems.

In the early days, researchers monitor the distribution and abundance of plankton with manually collecting plankton samples and classification by experts. The aforementioned process is so laborious and time consuming that hinders the plankton monitoring. To solve the problem, several imaging devices have been developed for collecting plankton images, and automatic classification system of plankton images also have been developed. The invention of imaging devices and classification system improves the efficiency of plankton monitoring. Meanwhile, the cost and specialized demand are reduced. At present, the generalization ability and classification performance wait for enhancing. This thesis studies the plankton image classification based on multiple kernel learning to improve the performance of classification system. So the system can be better used to monitor plankton. The main researches are as follows:

- Different plankton datasets are established for experiments. In order to reduce dataset bias and improve generalization ability, the plankton images gathered by different organizations and imaging devices are collected to build different datasets for experimental purposes. These datasets involve both phytoplankton dataset and zooplankton dataset.
- From various angles to analyze the characteristic of plankton, feasible and valid features can be combined and applied to describe the characteristic of plankton. During the process of research, firstly, according to the characteristic of plankton, statistical methods are used to extract geometric and gray features. In addition, classical algorithms in object detection are applied to describe characteristic of plankton, such as Local Binary Pattern, Histogram of Oriented Gradients, Inner-Distance Shape Context and so on. Then, the redundant information of extracted feature is reduced by feature selection to choose the best features for each dataset.
- The system based on multiple kernel learning is proposed for plankton classification.

In this process, predefined kernels are chose for each type of feature, and the optimal linear or non-linear combination of all kernels is learned with multiple kernel learning. So that features can give full play in classification.

- Contrast experiments are designed to establish system to evaluate the performance of plankton classification system. Firstly, according to the successful methods of plankton classification method at present, a baseline experiment is designed as the standard of contrast. Then, the second experiment is designed based on the baseline experiment to evaluate the property of feature extraction methods. Finally, we carry out the plankton classification system based on multiple kernel learning, and evaluate the performance of it by comparing with previous experiments.

Experimental results show that the plankton classification system based on multiple kernel learning has better classification performance and generalization ability. It is feasible and effective in plankton monitoring.

Keywords: **plankton, multiple kernel learning, feature extraction, kernel function, feature fusion**

目 录

1 绪论	1
1.1 课题研究的背景及意义	1
1.2 国内外研究现状	2
1.2.1 浮游生物图像采集技术	2
1.2.2 浮游生物图像分类技术	3
1.3 课题来源	4
1.4 论文内容和安排	4
2 浮游生物图像分类	6
2.1 浮游生物基本知识介绍	6
2.2 数据集介绍	7
2.2.1 WHOI 采集的数据集	7
2.2.2 ZooScan 系统采集的数据集	8
2.2.3 Kaggle 竞赛数据集	9
2.3 浮游生物图像分类方法介绍	10
2.3.1 伍兹霍尔海洋研究所浮游植物分类方法	10
2.3.2 法国国家科学院浮游动物图像扫描分析系统	11
2.4 评价方法介绍	12
2.4.1 混淆矩阵	13
2.4.2 F-Measure	14
2.4.3 交叉验证	14
2.5 本章小结	15
3 浮游生物特征分析	16
3.1 浮游生物特征提取	16
3.1.1 几何和灰度特征	16
3.1.2 粒子测度	19
3.1.3 纹理特征	20
3.1.4 局部特征	25
3.2 浮游生物特征选择	33
3.2.1 按搜索测量进行特征选择的方法	33
3.2.2 按评价准则进行特征选择的方法	34

3.3 本章小结.....	35
4 基于多核学习的浮游生物图像分类研究	36
4.1 多核学习理论	36
4.1.1 支持向量机	36
4.1.2 核函数	39
4.1.3 多核学习	41
4.2 基于多核学习的浮游生物图像分类系统.....	45
4.3 对比实验.....	48
4.3.1 基准实验	48
4.3.2 特征对比实验.....	49
4.3.3 基于多核学习的浮游生物图像分类实验	50
4.4 实验结果分析	53
4.5 本章小结.....	55
5 总结与展望	57
5.1 总结	57
5.2 展望	58
参考文献	59
附录 A 浮游生物种类	62
致 谢	64
个人简历、在学期间发表的学术论文与研究成果	65

1 绪论

1.1 课题研究的背景及意义

在我们所生活的地球上，海洋是其表面上最为广阔的水域，占 70% 以上。整个海洋环境以及生活在其中的动植物、微生物共同构成了海洋生态系统。在地球上，海洋生态系统为最大的生态系统，其中含有大量的浮游生物。浮游生物是海洋中其他生物的能量来源，因此为构成海洋生态系统必不可少的部分。浮游生物不仅包括浮游植物，还包括浮游动物。在海洋生态系统中，浮游植物是生成者，作为食物链的基本环节可以为其他的生物提供生活所需的能量。此外，浮游植物还会影响全球碳循环和海水中营养物质的浓度。而浮游动物以浮游植物为食，同时还是海洋食物链中高营养级动物的食物。因此，浮游动物作为海洋中生物间能力传递的桥梁，是海洋生态系统中必不可少的部分。综上，浮游动植物的丰富度及分布范围会影响整个海洋生态系统的平衡。

当浮游植物中有害的藻类大量繁殖时，会引发赤潮现象。赤潮会给海洋生态系统带来以下危害：(1) 赤潮藻大量聚集会使水体缺氧，鱼类容易窒息死亡；(2) 鱼类吞食有毒浮游植物会导致死亡；(3) 赤潮发生后会导致水体 pH 值升高，造成水中生物死亡。因此，浮游植物的大量繁殖将破坏生态平衡，影响渔业的发展，甚至危害人类健康。同时，浮游动物大量繁殖也会对海洋环境造成一定的影响：(1) 水域中浮游动物密度较高时会争夺海洋中其他生物的氧气；(2) 浮游动物作为海洋食物链中的重要环节，它的丰富程度会影响食物链的平衡；(3) 大量浮游动物聚集还会影响水下信号的传播。因此，人们越来越重视对海洋中浮游生物丰富程度的监测，这也是对海洋环境健康程度的一个评价指标。

浮游生物监测是统计在某一时间和空间范围内物种的数量以及丰富度。传统的浮游生物检测先采集样本，然后通过专业人员在显微镜下进行识别和统计，最终计算出该区域中浮游生物的丰富度。然而浮游生物体型小、数量多，人工进行采样、分类和统计不仅需要较高的专业水平，还需要消耗大量的人力、物力和时间。为了提高浮游生物监测的效率，人们研发出了浮游生物图像采集系统，可以方便的采集到水下的浮游生物图像。同时，利用图像处理和模式识别技术设计出浮游生物自动识别系统，对采集的浮游生物显微图像进行自动分类识别，可以实现对浮游生物的自动监测。这两个系统的出现大大提高了对浮游生物丰富度监测的效率，降低了成本。因此，目前人们越来越多的关注于研发性能更好的浮游生物自动分类系统，提高分类的准确率，扩大分类系统的适用范围，使其更好的应用于

浮游生物丰富度的监测。

1.2 国内外研究现状

浮游生物的自动监测主要包括浮游生物图像采集和分类识别两个部分，下面主要介绍这两部分的国内外研究现状。

1.2.1 浮游生物图像采集技术

传统的浮游生物采集通常使用网采、瓶采或泵采等方法，这些采集方法存在着一些问题：首先只能在相对较低的时间空间范围内采集样品，而且分析周期很长；其次，拖网容易扰乱浮游生物的分布结构^[1]。为了克服传统浮游生物采集方法的缺点，在过去的一段时间里浮游生物图像采集系统被广泛的研究与应用。浮游生物图像采集系统通常可以分为两类：实验室成像系统和原位图像采集系统。

实验室成像系统是指在实验室中使用的将浮游生物样本转换为数字图像的仪器设备。浮游动物图像扫描分析系统（ZooScan Integrated System）是一个比较有代表性的实验室成像系统，它由法国人 Gorsky. 等发明^[2]，主要用于对采集到的液体样本中的浮游动物进行成像、检测、识别，该设备由 ZooScan、ZooProcess 和 Plankton Identifier 三部分组成。其中 ZooScan 是扫描成像部分，即图像采集部分，主要负责将采集到的浮游动物样本通过扫描的方式转换成数字图像。ZooProcess 和 Plankton Identifier 主要对 ZooScan 得到的图像进行处理、测量和自动分类。目前，浮游动物图像扫描分析系统已经广泛应用于浮游动物图像采集和自动分类识别并有较高的效率和分类准确率，在国际上被广泛认可并投入商业化生产^[3]。

浮游生物原位图像采集系统可以实时采集水下浮游生物原位图像，保证浮游生物的生存分布结构不被破坏。早在 1992 年 Davis 等人研制出了浮游生物视频记录器（Video Plankton Recorder, VPR）^[4]，这是最早的用来采集浮游生物原位图像的系统。后来随着浮游生物原位图像采集系统不断发展，先后出现了水下视频剖面仪（Underwater Video Profiler, UVP）^[5]、流式细胞仪（Flow Cytometer and Microsocpe, FlowCAM）^[6]、灰度图像颗粒探测系统（Shadowed Image Particle Platform and Evaluation Recorder, SIPPER）^[7]、流式成像技术（Imaging FlowCytobot, IFCB）^[8] 等设备。这些设备的出现大大提高了采集浮游生物图像的效率，方便了对水下浮游生物丰富度的监测。

1.2.2 浮游生物图像分类技术

传统的浮游生物分类方法是生物学家根据其掌握的专业知识对采集的浮游生物样本进行人工分类。然而海洋中浮游生物种类繁多、形态各异，这使得传统人工分类方法存在以下几个问题：首先，对浮游生物进行分类时需要人员具有较高的专业水平；其次，浮游生物个体小、数量多，人工分类不仅需要大量人力，还会消耗大量时间；第三，人工分类速度较慢，难以实现对浮游生物丰富度的实时监测。因此，人们结合浮游生物图像采集系统研究出了浮游生物自动分类系统。

浮游生物自动分类系统通常采用图像处理和模式识别算法，可以对图像采集设备收集的浮游生物图像进行快速自动分类识别。其中，图像处理是为了获得图像中有用的信息，采用计算机对采集到的数字图像进行去噪、分割、增强并提取特征等操作。

早在 20 世纪末，硅藻图像数据库已经建立，在对该图像进行自动识别过程中人们结合了图像处理和模式识别方法。在 1996 年 Culverhouse 研发了一个对甲藻进行分类的系统，该系统提取图像中细胞的形状和表面特征进行分析，并采用了人工神经网络方法进行分类，实验结果在 3 种甲藻图像上的总体分类准确率可达 72%。汤晓鸥在 1998 年^[9] 提出采用不变矩和傅里叶描述子对浮游生物视频记录器采集的浮游生物图像进行分类，该方法对含有接近 2000 张图像的浮游生物数据集（包括 6 类浮游生物）进行分类，得到的分类结果可以达到 95%。后来，在 2005^[10] 和 2006^[11] 两年中，汤晓鸥还提出采用形状特征对二值浮游生物图像进行分类。Hu 等人提出使用灰度共生矩阵描述图像中目标的灰度特征，然后采用支持向量机训练分类器。在 2007 年 Sosik 等人^[12] 结合多种特征设计了一个浮游生物分类系统，这些特征包括大小、形状、对称性、纹理等，并使用特征选择算法去掉其中的冗余部分，然后针对选择的结果采用支持向量机训练分类器，该分类系统在 22 类浮游生物图像上的分类准确率达到 88%。之前提到的 ZooScan Integrated System^[13] 中的 Plankton Identifier 是对 ZooScan 采集的浮游生物图像进行分类和识别的部分，该部分主要根据系统中提取的一系列描述浮游生物的形状、灰度等标量（例如面积、周长、圆形度、灰度对比度、曲率等）来进行分类，该系统在含有 20 类浮游动物的不平衡数据集上可以得到约 78% 的分类准确率。Mosleh 等^[14] 采用形状和纹理特征对藻类进行描述，然后通过神经网络进行识别。在 2015 年 Ellen 等人^[15] 从不同角度对浮游生物分类方法进行分析，研究如何提高浮游生物分类的准确率。

分析国内外的研究现状可以发现，目前的浮游生物自动分类系统已经可以高效、准确的实现对采集的图像进行分类，然而也存在着一些有待改进的方面：（1）现有的分类系统中使用的特征提取方法较为单一，并不能全面的描述图像中浮游

生物的形态特征；(2) 部分系统在分类过程中使用多种特征描述浮游生物的形态特征，然而在融合不同特征时没有考虑每种特征的贡献比例，并不能充分发挥每种特征的积极作用；(3) 由于浮游动物个体相对于浮游植物较大，并且形态相对复杂，因此人们设计的分类系统大多不能同时适用于浮游植物和浮游动物；(4) 由于浮游生物种类繁多，国内外目前设计的自动分类系统大多只针对几个特定类别的浮游生物，适用范围较窄。

因此根据以上问题，我们结合多核学习设计了一个浮游生物图像分类系统，该系统从不同角度提取了浮游生物的多种特征，对形态进行全面描述；然后使用多核学习算法将所有特征融合，充分发挥每种特征在分类过程中的积极作用，提高系统的分类性能。同时，该系统具有广泛的适用范围，可以应用于不同的浮游生物数据集（既包括浮游植物也包括浮游动物），具有较好的泛化能力。

1.3 课题来源

国家自然科学基金项目“基于视觉注意结合生物形态特征的海洋浮游植物显微图像分析”(批准号：61301240)、国家自然科学基金项目“基于生物形态特征的中国海常见有害赤潮藻显微图像识别”(批准号：61271406)、中央高校基本科研业务费项目“海洋浮游动物原位探测与分析系统”(批准号：201562023)。

1.4 论文内容和安排

本文的主要工作内容和安排如下：

第一部分为绪论，该部分主要针对浮游生物分类研究的背景意义、国内外研究现状以及课题来源进行介绍。

第二部分介绍浮游生物图像分类的预备知识，包括浮游生物的基本知识、后续实验使用的数据集、目前应用较广的浮游生物图像分类方法以及评价浮游生物分类系统性能所使用的评价方法。

第三部分根据对浮游生物形态特征的分析以及人们对浮游生物分类识别的过程，选用适合的特征提取方法，既包括简单的几何灰度特征，也包含计算机视觉领域的经典算法，例如局部二值模式、内距离形状上下文等。然后介绍特征选择算法，该方法可以去除提取特征中存在的冗余信息，保留最有效部分，从而降低特征维数，提高分类器的性能。

第四部分首先以支持向量和核函数理论为基础，介绍多核学习的基本思想，以及简单多核学习和非线性多核学习方法。然后，对设计的基于多核学习的浮游生

物图像分类系统进行介绍。最后，设计一系列对系统性能进行评价的对比实验，根据对比实验的结果对基于多核学习的浮游生物图像分类系统的性能进行分析。

第五部分对本文研究的浮游生物图像分类工作进行总结，同时针对研究过程中存在的问题进行分析和展望。

2 浮游生物图像分类

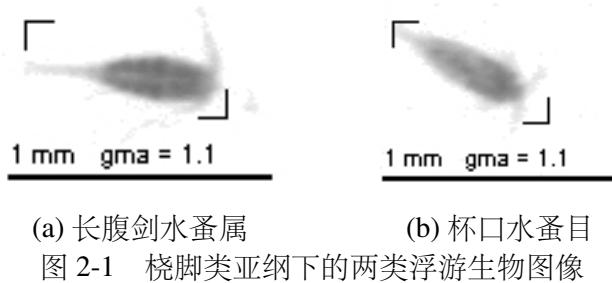
研究浮游生物图像自动分类，既需要了解浮游生物领域的相关知识，也要掌握计算机视觉和机器学习领域的相关理论。本章作为介绍基于多核学习的浮游生物图像分类研究的预备知识章节，首先介绍浮游生物的基础知识，然后对后续实验使用的数据集进行说明，接下来介绍目前分类性能较好的浮游生物图像分类方法，在本章最后介绍评价分类系统性能所采用的评价方法。

2.1 浮游生物基本知识介绍

浮游生物是指生活在水中运动能力较弱的生物体，它们是许多大型水生生物的食物，包括浮游植物和动物两大类。浮游植物是一种微小的植物，在水中以浮游状态生存，通常是指浮游藻类，包括裸藻门、黄藻门、金藻门、绿藻门、甲藻门、硅藻门、蓝藻门、隐藻门八个门类，目前已知的浮游植物约有4万种。浮游动物是生活在水中的无脊椎动物和脊索动物幼体的总称，其种类繁多，主要的门类有原生动物、浮游幼虫、甲壳纲、毛颚动物、腔肠动物、被囊动物等。

浮游生物通常较小，个体从几微米到几毫米大小不等，按照个体的大小可以将其分为以下几类：小于5微米为超微型浮游生物；5至50微米之间为微型浮游生物；50微米到1毫米之间为小型浮游生物；1至5毫米之间为中型浮游生物；5毫米到10毫米间为大型浮游生物；大于1厘米为巨型浮游生物。通常浮游植物的个体相对浮游动物较小，一般属于微型和小型浮游生物；而浮游动物体型通常较大，主要为中型、大型以及巨型浮游生物。由于大多浮游生物较小，因此必须使用显微镜进行观测。

浮游生物是按照界、门、纲、目、科、属、种的分类学原理进行分类的，从上层的“界”到下层的“种”，越往下层被归为同一个分支的浮游生物之间形态特征越相似。因此在对浮游生物进行分类时，时常会遇到形态特征十分相似的两个不同类别的浮游生物生物，这一特点增加了浮游生物分类识别的难度。例如，图2-1(a)中的浮游生物为长腹剑水蚤属，而图2-1(b)中为杯口水蚤目，这两幅图像中的浮游生物分别属于桡脚类亚纲下的剑水蚤目和杯口水蚤目，但它们的形态特征十分相似，不易区分。



(a) 长腹剑水蚤属
(b) 杯口水蚤目
图 2-1 桡脚类亚纲下的两类浮游生物图像

2.2 数据集介绍

为了设计一个泛化能力强、适用范围广的浮游生物图像分类系统，本文搜集构建了以下三个不同浮游生物数据集进行实验：一个是由伍兹霍尔海洋研究所（Woods Hole Oceanographic Institution, WHOI）使用 FlowCytobot 采集的浮游植物数据集；另一个是使用 ZooScan 系统采集的浮游动物数据集；还有 Kaggle 竞赛中使用的浮游生物数据集。

2.2.1 WHOI 采集的数据集

在美国的大西洋海岸上有一个综合性海洋科学研究院——伍兹霍尔海洋研究所，其致力于对海洋中各个领域进行研究。本数据集^[12]是由该机构研究人员使用 FlowCytobot 采集的 2004 至 2005 年间伍兹霍尔港附近的浮游植物图像组成，该数据集一共包括 22 类浮游植物图像，其例图如图 2-2 所示。

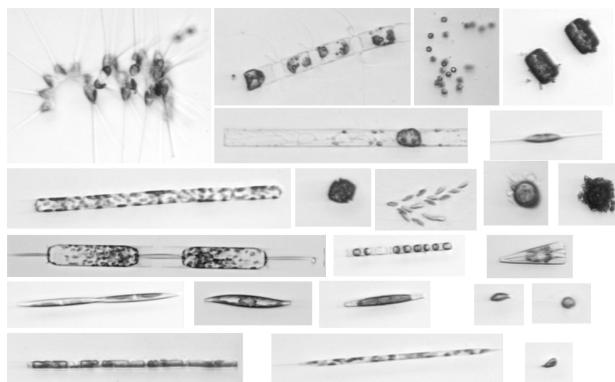


图 2-2 WHOI 采集的数据集中每类浮游植物的例图

这 22 类图像中包含 16 类硅藻类图像：(1) 星杆藻属 (*Asterionellopsis*), (2) 角毛藻属 (*Chaetoceros*), (3) 细柱藻属 (*Cylindrotheca*), (4) 指管藻属 (*Dactyliosolen*), (5) *DactfragCeratul*, (6) 双尾藻属 (*Ditylum*), (7) 几内亚藻属 (*Guinardia*), (8) 楔形藻属 (*Licmophora*), (9) 斜纹藻属 (*Pleurosigma*), (10) 拟菱形藻 (*Pseudonitzschia*), (11) 根管藻属 (*Rhizosolenia*), (12) 骨条藻属 (*Skeletonema*), (13) 海链藻属 (*Thalassiosira*), (14)

锥囊藻属 (Dinobryon), (15) 眼虫属 (Euglena), (16) 棕囊藻属 (Phaeocystis)。除此之外还有 4 类由形态相似的浮游生物图像构成: (1) 各种形状的纤毛虫 (ciliate), (2) 双鞭毛虫门 (dinoflagellate), (3) 鞭毛虫 (nanoflagellate), (4) 有翼的硅藻类 (pennate diatoms)。另外还有两种海洋中的其他物质: 一种是个体小于 $20\mu m$ 的不明物; 另一种是碎石。整个数据集分为训练集和测试集, 各包含 3300 张浮游植物图像, 共 6600 张, 其中每个类别中图像数量相等。

2.2.2 ZooScan 系统采集的数据集

该数据集^[13]由 ZooScan 系统^[2]采集的浮游动物图像组成。数据集中包含 20 类浮游动物图像, 其例图如图 2-3 所示。这 20 类图像中有 14 类为浮游动物: (1) 蠕螺属 (Limacina), (2) 翼足目 (Pteropoda), (3) 尖头溞属 (Penilia), (4) 长腹剑水蚤属 (Oithona), (5) 杯口水蚤目 (Poecilostomatoidea), (6) 桡脚类亚纲 (Copepoda) 中的其他类生物, (7) 十足目 (Decapoda), (8) 尾海鞘纲 (Appendicularia), (9) 樽海鞘纲 (Thaliaceae), (10) 毛鄂动物门 (Chaetognatha), (11) 各类浮游生物的卵, (12) 放射虫门 (Radiolaria), (13) 钟泳亚目 (Calycophorae), (14) 水母亚门 (Medusae)。另外 6 类为非浮游生物: (1) 气泡 (bubble), (2) 纤维 (fiber), (3) 聚集物 (aggregates), (4) 暗色聚集物 (dark aggregates), (5) 假浮游生物 (pseudoplankton), (6) 采集到的聚焦不好的图像。该数据集共 3771 张图像, 每类的图像数量各不相同, 图 2-4 显示了数据集中每个类图像的数量, 数量最少的类别有 28 张图像, 最多的类别有 427 张。

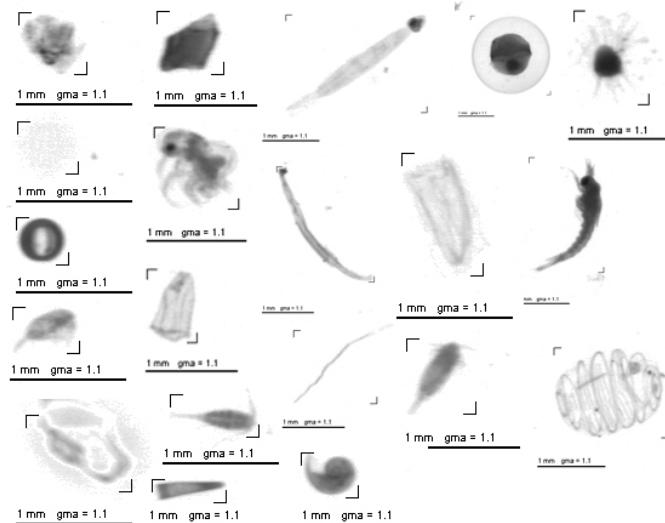


图 2-3 ZooScan 系统采集的数据集中每类浮游动物的例图

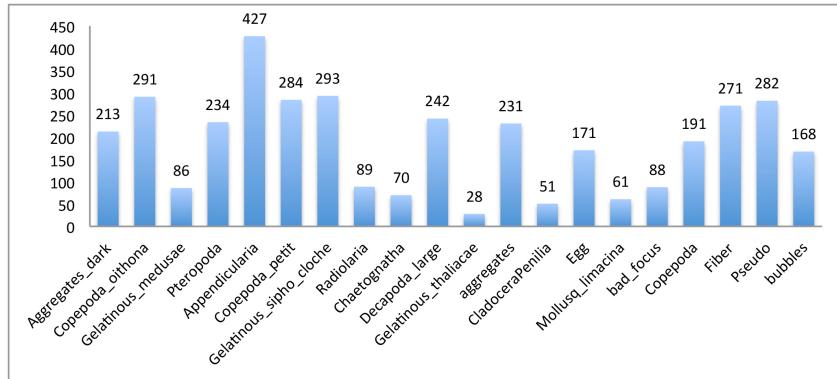


图 2-4 ZooScan 系统采集的数据集中每类图像的数量

2.2.3 Kaggle 竞赛数据集

Kaggle 是一个数据分析的竞赛平台，研究者或企业可以将数据、问题发布在 Kaggle 平台上，通过竞赛的方式向大家征集解决方案。为了预测海洋健康程度，并为促进海洋健康做贡献，Kaggle 竞赛平台上组织了浮游生物识别竞赛，根据浮游生物种群的丰富度衡量海洋生态系统的健康程度。Kaggle 平台上的竞赛数据集由俄亥俄州立大学菲尔德海洋科学中心采集并提供，训练集共 121 类。本文实验中采用的数据集为该竞赛训练集的一部分，共选用 38 类浮游生物，例图如图 2-5 所示，其中 35 类为浮游生物，另外 3 类为非浮游生物。该数据集共 28748 张图像，每个类别数量各不相同，图 2-6 显示了每类图像的数量，数量最少的类别仅有 108 张图像，最多的有 1979 张图像。

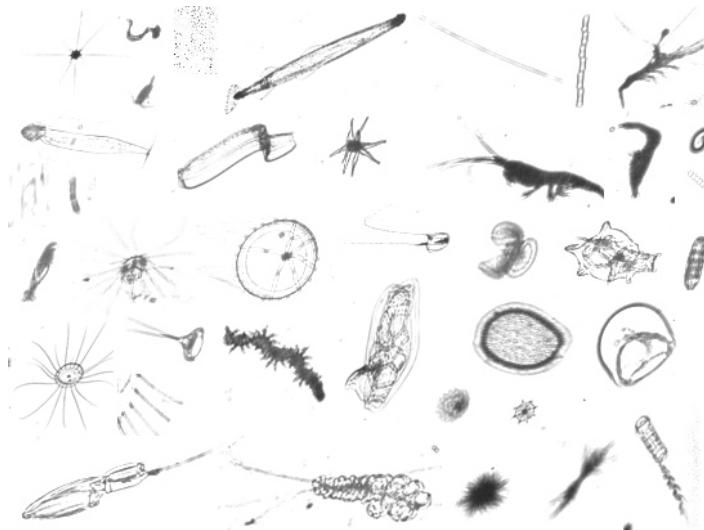


图 2-5 Kaggle 竞赛数据集中每类浮游生物的例图

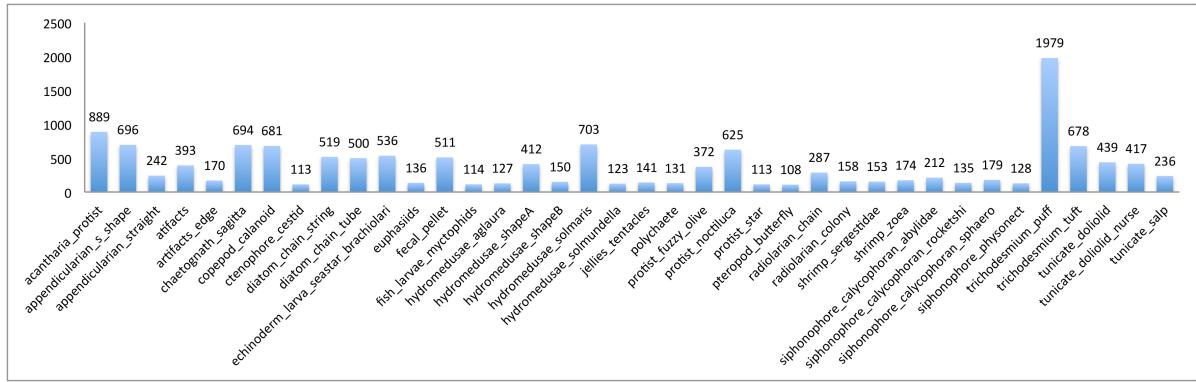


图 2-6 Kaggle 竞赛数据集中每类图像的数量

2.3 浮游生物图像分类方法介绍

目前对浮游生物图像分类的研究众多，下面分别介绍两种浮游生物图像分类系统：一个为浮游植物自动分类，由伍兹霍尔海洋研究所的 Sosik 等人提出；另一个是由法国国家科学院 Gorsky 等人研制的浮游动物图像扫描分析系统。这两个系统不仅具有较好的分类性能，而且被广泛应用于浮游生物分类，在后续实验中根据它们可以设计浮游生物分类系统性能的评价基准。

2.3.1 伍兹霍尔海洋研究所浮游植物分类方法

由于浮游生物个体小并且数量多，人工对其进行分类是不切实际的。伍兹霍尔海洋研究所的研究人员为了处理 FlowCytobot 采集的大量浮游植物图像，他们根据特征提取和模式识别算法提出一种浮游植物自动分类系统，该系统主要包括以下几个部分：

1. 图像预处理。在对图像中浮游植物形态特征进行提取时，除了需要目标的灰度纹理信息外，还需要用到目标的轮廓信息，因此要进行图像预处理，提取图像中目标生物的轮廓。在提取目标轮廓时，先采用相位一致性计算、边缘检测、数学形态学处理得到目标生物的二值图像，然后重新提取目标区域的简单边缘，从而可以获得图像中浮游生物的轮廓信息。
2. 特征提取。根据采集的原始浮游生物图像以及其对应的二值图像和轮廓边缘可以提取以下特征：简单的几何特征，例如目标的周长、长宽比、面积等；形状和对称性特征；纹理特征；不变矩；灰度共生矩阵等，共得到 210 个特征信息。这些特征主要通过 MATLAB、DIPUM 工具箱以及自定义函数生成。
3. 特征选择。由于采集的浮游生物特征中可能包含冗余或不相关的信息影响分类器的整体性能，因此采用特征选择去除冗余特征。
4. 训练分类器。特征选择后采用支持向量机来训练分类器，使用的核函数为高

斯核函数（即径向基核函数），并采用 10 折交叉验证来确定核函数中的最优参数，该部分实验主要使用 LIBSVM 函数库实现。

该分类方法在 2.2.1 数据集上进行实验，使用训练集训练分类器，用获得的分类器对测试集图像进行分类，得到分类准确率可以达到 88%，其中有 12 个类别的分类结果超过了 90%，只有 4 类图像低于 80%。

2.3.2 法国国家科学院浮游动物图像扫描分析系统

浮游动物图像扫描分析系统（ZooScan Integrated System）是由法国国家科学院 Villefranche 海洋实验 Gorsky 等人发明的实验室浮游动物成像系统，该系统不仅可以完成对水下浮游动物从采集，而且能够对采集的图像进行识别^[3]。与其他设备相比，ZooScan 系统方便实用，并且具有较好的分类性能，目前该系统已经被广泛的商业化生产。

Zooscan 系统由三部分组成，分别是 ZooScan、ZooProcess 和 Plankton Identifier (PkID)。其中 ZooScan 为图像采集部分，将采集的浮游动物样本放在 ZooScan 仪器上，通过扫描形成数字图像。ZooProcess 为软件部分，对 ZooScan 采集的浮游动物图像进行处理，分割提取并测量图像中的个体，可以自动得到图像中每个目标的参数，如灰度、大小、形状等参数。Plankton Identifier 可以根据得到的参数对浮游生物图像进行分类。用 ZooScan 系统采集并识别浮游动物图像的过程如图 2-7 所示。

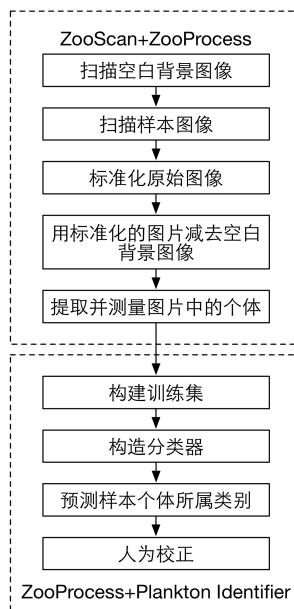


图 2-7 ZooScan 系统的基本流程

使用 ZooScan 系统对采集的浮游动物进行分类时，首先用 ZooProcess 软件控制 ZooScan 扫描仪采集浮游动物图像：（1）扫描空白的背景图像；（2）扫描样本图像；（3）标准化原始图像；（4）用标准化得到的样本图像减去背景，并去除图像中的扫描框；（5）提取图像中的个体并测量每个个体的特征参数，测量得到的浮游生物特征参数共 67 个，大致可以分为四类：

1. 灰度参数，例如：

- Max, Min 物体灰度的最大、小值
- Mean 物体灰度的平均值
- IntDen 物体灰度值的总和
- StdDev 物体灰度值的标准差

2. 大小参数，例如：

- Area 物体面积
- Perim 物体的周长
- Feret 物体的最大费雷特径

3. 形状参数，例如：

- Fractal 物体的分形维数
- Circ 物体的圆形度
- Skelarea 物体骨架的像素数量

4. 位置参数，例如：

- X, Y 物体的重心
- XM, YM 物体的灰度重心
- Height, Width 物体最小外接矩形的高和宽

然后，根据 ZooProcess 提取的特征参数，采用 Plankton Identifier 对采集的浮游动物图像进行分类：（1）构建训练集；（2）选用模式识别算法来训练分类器，PkID 中提供的算法有 K 近邻、随机森林、C4.5、多层次感知机、支持向量机（线性核函数、高斯核函数）等；（3）用训练得到分类器可以识别未知的浮游生物图像；（4）对预测结果进行人工校准。该分类系统在其采集的 2.2.2 数据集上的分类准确率可以达到 78%，分类效果最好的类别其准确率可以达到 90% 以上。

2.4 评价方法介绍

评价一个分类系统性能主要从分类的准确度和可靠性两个方面进行，常用的评价方法有混淆矩阵、ROC 曲线等，本文中采用的是混淆矩阵。

2.4.1 混淆矩阵

在机器学习领域，混淆矩阵（Confusion Matrix, CM）也被称为误差矩阵，是用来呈现算法性能的可视化工具，通常用于监督学习。若有 n 个类别，则混淆矩阵由 n 行 n 列组成，其中每一列表示每一类的预测样本数量，每一行表示每一类的实际样本数量，识别正确的样本数量为对角线，如表 2-1 所示。

表 2-1 混淆矩阵

		预测结果	
		正样本	负样本
实际结果	正样本	a	b
	负样本	c	d

上表中 a 表示真阳性（True positives, TP），即正样本预测正确的数量； b 表示正样本预测错误的样本数，即假阴性（False negatives, FN）； c 表示负样本被分为正样本的数目，即假阳性（False positives, FP）； d 表示负样本预测正确的数量，即真阴性（True negatives, TN）。根据这些数据可以计算出以下几个指标：

- 真阳性率（True positive rate, TPR），也是召回率（Recall），它表示正样本被正确识别的概率。

$$TPR = \frac{a}{a + b} \quad (2-1)$$

- 假阳性率（False positive rate, FPR）表示负样本被误分为正样本的概率。

$$FPR = \frac{c}{c + d} \quad (2-2)$$

- 真阴性率（True negative rate, TNR）表示负样本被正确分类的概率。

$$TNR = \frac{d}{c + d} \quad (2-3)$$

- 假阴性率（False negative rate, FNR）表示正样本被误分为负样本的概率。

$$FNR = \frac{b}{a + b} \quad (2-4)$$

- 错误发现率 (False discovery rate, FDR) 表示被预测为正样本中负样本的概率。

$$FDR = \frac{c}{a + c} \quad (2-5)$$

- 阳性预测值 (Positive predictive value, PPV)，也称为命中率 (Precision)，表示在预测为正样本的样本中真正的正样本所占的比重。

$$Precision = \frac{a}{a + c} \quad (2-6)$$

本文实验中主要采用以下两个指标: Recall 和 Precision。由于 Recall 和 Precision 有时会出现矛盾情况，因此本文采用 F-Measure 评价分类器的性能。

2.4.2 F-Measure

F-Measure 是一种综合评价指标，也被称为 F-Score，当 Recall 和 Precision 出现矛盾时，就可以采用该方法进行评价。

$$F = \frac{(\alpha^2 + 1)P * R}{\alpha^2(P + R)} \quad (2-7)$$

其中 α 为参数， P 代表 Precision， R 为 Recall。

当 $\alpha = 1$ 就得到 F1-Measure：

$$F1 = \frac{2 * PR}{P + R} \quad (2-8)$$

F1 的值较高说明试验分类模型性能更好。

2.4.3 交叉验证

交叉验证 (Cross Validation) 是一种检验分类器性能的方法，该方法将数据集分为训练集和验证集，首先用训练集训练分类器，然后用获得的分类器对验证集进行识别，其分类结果为评价分类器性能的指标^[16]。在评价过程中使用交叉验证可以得到更加稳定可靠的评价结果。常用的交叉验证形式有：Hold-out 验证、K 折交叉验证 (K-fold Cross Validation, K-CV)、留一验证 (Leave-One-Out Cross Validation, LOO-CV)。

- Hold-out 验证是将数据集划分为两集合：一个作为训练集，用其训练分类器；另一个作为验证集，用其对分类器进行测试，最终得到的分类结果可以作为该分类器的性能指标。
- K 折交叉验证就是将数据集划分为 K 个子集，将每一个子集分别看做验证集进行实验，与此同时另外 $K-1$ 个部分作为训练集。一共要进行 K 次实验，训练获得 K 个分类模型，每个模型对验证集进行分类会得到一个准确率，这 K 个准确率的平均值可以作为该分类器评价指标。
- 留一验证的基本思想为：若原数据集一共 n 个样本，则将每个样本单独看做验证集，另外 $n-1$ 个组成训练集。与 K 折交叉验证类似，留意验证一共需要进行 n 次实验，获得 n 个训练模型和分类准确率，计算 n 个准确率的平均值可以作为分类器的评价指标。

本文对分类器性能进行评价时采用 K 折交叉验证。

2.5 本章小结

在介绍基于多核学习的浮游生物分类研究之前，本章先介绍了浮游生物分类的相关背景。首先介绍的是浮游生物的相关内容；然后详细介绍后续实验中所用的数据集；接下来介绍目前分类性能较好的浮游生物分类方法，在后面的实验里将根据它们设计对比实验的基准；在本章最后介绍了对分类系统性能进行评价的方法，本文使用了 K 折交叉验和混淆矩阵来统计得到的分类结果，从而对分类器性能进行评价。

3 浮游生物特征分析

在浮游生物分类过程中，特征分析是一个重要环节，得到特征的好坏会直接影响分类的最终结果。我们对浮游生物特征进行分析，根据浮游生物形态特征，从各个角度选取适合的特征提取方法进行特征描述，然后从获得的特征中选取有效的特征子集。本章主要介绍对浮游生物形态特征进行提取的方法，以及去除特征中冗余信息采用的特征选择算法。

3.1 浮游生物特征提取

人们在识别浮游生物过程中，会根据浮游生物的形状、纹理等特征对浮游生物进区分。浮游生物自动分类识别系统正是模拟人类对浮游生物分类识别的过程进行设计的，因此选择特征提取方法时可以结合人类在识别浮游生物时采用的特征。在本文中，根据浮游生物的形态特征，并结合经典计算机视觉算法对浮游生物特征进行提取，主要使用的方法有以下几种：几何灰度特征；粒子测度；纹理特征，包括变差函数、Gabor 滤波器、二元梯度轮廓、局部二值模式；局部特征，包括方向梯度直方图、内距离形状上下文、尺度不变特征变换。

3.1.1 几何和灰度特征

目标的几何灰度为浮游生物的基本特征，简单的几何和灰度（例如面积、周长、灰度对比度等）等统计标量可以用来表示浮游生物的形态特征。在本文的研究中，共使用 43 个几何和灰度特征对浮游生物进行描述，下面介绍其中部分特征。

几何标量特征：

1. 周长：为图像中包围目标的轮廓边缘上所有像素数量之和。

$$L = \sum_{x=1}^M \sum_{y=1}^N n(x, y) \quad (3-1)$$

其中， L 表示周长， x, y 为像素点坐标， $n(x, y)$ 表示物体的边缘二值图， M, N 为图像高和宽。

2. 面积：是指图像中目标区域内像素数量之和。

$$A = \sum_{x=1}^M \sum_{y=1}^N f(x, y) \quad (3-2)$$

其中， A 为面积， $f(x, y)$ 代表目标的二值图。

3. 体态比：表示可以包围目标的最小外接矩形的长和宽之比。

$$C = \frac{W}{H} \quad (3-3)$$

其中， C 为体态比， W, H 为最小外接矩形的长宽。

4. 圆形度：表示目标区域的圆形程度。

$$e = \frac{4\pi A}{L \times L} \quad (3-4)$$

其中， e 表示圆形度，当 e 等于 1 时目标的形状为圆形， e 越小表示目标的形状越不规律，形状与圆形差距较大。

5. 伸长率：是指可以拟合目标形状的最优椭圆的长轴和短轴之比。

$$\delta = \frac{\text{Major}}{\text{Minor}} \quad (3-5)$$

其中， δ 表示伸长率， Major 表示椭圆的长轴， Minor 表示短轴。

6. 凸率：是指目标物体的面积与目标物体的凸壳面积之间的比值。

$$C_R = \frac{A}{\sum_{x=1}^M \sum_{y=1}^N k(x, y)} \quad (3-6)$$

其中， C_R 为凸率， $k(x, y) = \begin{cases} 1 & (x, y) \in \text{目标区域凸包部分} \\ 0 & (x, y) \notin \text{目标区域凸包部分} \end{cases}$ 。

7. 等效球直径：是指与目标物体体积相同的球的直径。

$$Esc = 2 \times \sqrt{\frac{A}{\pi}} \quad (3-7)$$

8. 最大费雷特直径：弗雷特直径是一种粒径表示方法，沿着一定方向可以测量获得的投影目标区域轮廓两边界平行线间的距离。最大弗雷特直径即目标物体轮廓边界上两平行线间距离的最大值。

灰度标量特征：

1. 最小值：是指目标范围内全部像素点的最小灰度值。

$$I_{min} = \min_I I(x, y) \quad (3-8)$$

其中, I 表示灰度图像, x, y 表示图像中像素点的坐标。

2. 最大值: 是指目标范围内全部像素点的最大灰度值。

$$I_{max} = \max_I I(x, y) \quad (3-9)$$

3. 平均值: 是指目标范围内全部像素点灰度值的均值。

$$I_{avg} = \frac{\sum_I I(x, y)}{A} \quad (3-10)$$

4. 总密度: 是指目标范围内所有像素点灰度值之和。

$$I_{den} = \sum_I I(x, y) \quad (3-11)$$

5. 标准差: 是指目标范围内灰度值的标准差。

$$I_\sigma = \sqrt{\frac{1}{A} \sum_I (I(x, y) - I_{avg})^2} \quad (3-12)$$

6. 对比度: 表示目标区域中最亮和最暗像素区域的对比程度。

$$CON = \sum_{x=1}^k \sum_{y=1}^k G(x, y)^2 \quad (3-13)$$

其中, G 表示目标图像的灰度共生矩阵, k 表示目标图像中灰度值级数。

7. 自相关: 反映目标区域中纹理的一致性。

$$COR = \sum_{x=1}^k \sum_{y=1}^k \frac{xyG(x, y) - u_x u_y}{s_x s_y} \quad (3-14)$$

其中

$$u_x = \sum_{x=1}^k \sum_{y=1}^k xG(x, y), \quad u_y = \sum_{x=1}^k \sum_{y=1}^k yG(x, y) \quad (3-15)$$

$$s_x^2 = \sum_{x=1}^k \sum_{y=1}^k G(x, y)(x - u_y)^2, \quad s_y^2 = \sum_{x=1}^k \sum_{y=1}^k G(x, y)(y - u_y)^2 \quad (3-16)$$

8. 熵：是图像中所具有信息量的度量。

$$ENT = - \sum_{x=1}^k \sum_{y=1}^k G(x, y) \log G(x, y) \quad (3-17)$$

除了上述几种几何和灰度特征以外，在本文还使用了不变矩、对称性等标量值来描述浮游生物的基本形态特征。

3.1.2 粒子测度

粒子测度（Granulometry）由 Matheron 等人在 1978 年提出^[17]，可以用来计算二值图像中目标区域的大小分布情况。粒子测度的基本思想是对二值图做开运算，在开运算过程中不断增加结构元素的大小，并记录下在此过程中图像中区域内像素数量的变化。

$$G \circ T = \cup\{T + x : T + x \subset G\} \quad (3-18)$$

$$\psi_\lambda(G) = G \circ \lambda T \quad (3-19)$$

其中 G 为二值图， T 为结构元素， \circ 为开运算， λ 为一个正数变量，表示进行开运算的次数， $\psi_\lambda(G)$ 表示经过开运算后得到的二值图像。

$$F_G(\lambda) = 1 - \frac{v(\psi_\lambda(G))}{v(G)} \quad (3-20)$$

其中 $v(G)$ 表示得到原始图像中目标区域的像素数量， $F_G(\lambda)$ 为粒子分布，它可以描述图像中目标的区域分布特征。

在本文采用粒子测度提取浮游生物的特征，在实验中设置了两组参数，分别采用不同变换的结构元素进行开运算。第一组参数是将结构元素大小由 2 变到 50，间隔为 4，记录下图像中目标区域像素数量的变化情况。第二组参数是将结构元素大小由 5 变到 60，间隔为 5，记录图像中像素数量的变化。

3.1.3 纹理特征

3.1.3.1 变差函数

变差函数（Variogram）是 Motheron 在 1965 年提出的一种矩估计方法，可以表示区域化变量的随机性，反应了区域化变量在某个方向上某一距离范围内的相关程度。从描述纹理特征的角度来考虑，图像中浮游生物的灰度值可以看成反应图像纹理的随机性和相关程度的区域化变量^[18]，因此用变差函数能够有效表示图像的纹理特性，具体形式如下：

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x) - Z(x + h)]^2 \quad (3-21)$$

其中 h 是两个像素之间沿着某一方向的距离， $N(h)$ 是指在一定的区域内像素间距离为 h 的对数， $Z(x)$ 和 $Z(x + h)$ 为点 x 和 $x + h$ 的灰度值。用变差函数描述纹理特征时，选用一定的窗口、步长和方向计算函数值，然后取平均值赋给窗口中心点，得到差变函数纹理图。

3.1.3.2 Gabor 滤波器

Gabor 滤波器是一种基于信号处理的纹理分析方法，其实质是一种加了高斯窗口的傅里叶变换，通过窗口的变化可以提取图像不同尺度和不同位置上的纹理信息。Gabor 变换符合人和动物的视觉机理，具有良好的方向选择性和空间局部性^[19]，可以精确的描述局部纹理特性，因此该方法被广泛的用来提取纹理特征。本文中使用二维 Gabor 滤波器来描述浮游生物的纹理特征。

二维 Gabor 滤波器的冲击响应为^[20]：

$$h(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2} + 2\pi jF(x \cos \theta + y \sin \theta)} \quad (3-22)$$

其中 θ 为方向， F 为中心频率，这两个参数决定着 Gabor 滤波器的位置，即改变这两个参数可以得到不同滤波通道。Gabor 滤波器的本质就是与原图像进行卷积，即

$$Q(x, y) = I(x, y) * h(x, y) \quad (3-23)$$

其中 $I(x, y)$ 为原图像， $Q(x, y)$ 为 Gabor 滤波的结果。采用不同参数卷积后会得到不同的输出图像，每幅输出图像的均值和标准差可以表示该图像的纹理特性，所

有图像的值可以组成一个特征向量来描述原始图像的纹理特征。

$$\text{mean} = \frac{\sum_{x=0}^{n-1} \sum_{y=0}^{m-1} Q(x, y)}{m \times n} \quad (3-24)$$

$$\text{std} = \sqrt{\frac{\sum_{x=0}^{n-1} \sum_{y=0}^{m-1} [Q(x, y) - \text{mean}]^2}{m \times n}} \quad (3-25)$$

其中 m, n 表示图像的长和宽, mean 和 std 分别为卷积后图像的均值和标准差。若在实验中参数选用 a 个方向, b 个中心频率, 那么一共会得到 $a \times b$ 个不同的 Gabor 滤波器, 从而获得 $a \times b$ 张图像。然后分别计算每张滤波后得到图像的均值和标准差, 将获得 $a \times b \times 2$ 个值组成一个特征向量, 该向量可以有效的描述目标图像的纹理特征。在本文实验中, 我们为 Gabor 滤波器选定了不同的参数, 包括 6 个频率、8 个方向, 如图 3-1。用 Gabor 滤波器提取特征时, 每幅图像会获得一个 96 ($6 \times 8 \times 2$) 维的特征向量。

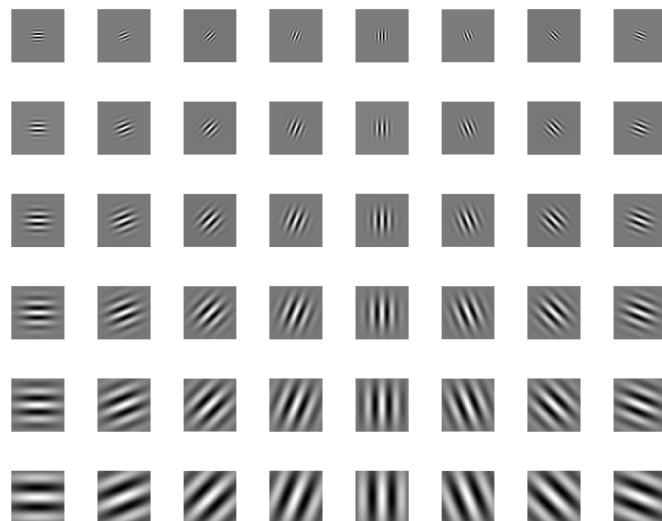


图 3-1 不同参数下的 Gabor 滤波器

3.1.3.3 局部二值模式

局部二值模式 (Local Binary Pattern, LBP) 由 Timo Ojala 等人在 1994 年提出^[21], 它是一种描述图像纹理特征的方法, 该算法有灰度不变性和旋转不变性。局部二值模式的基本思想是用 LBP 值描述目标中每一个像素与其邻域像素之间的差别, 然后通过统计整幅图像中的 LBP 值出现的频率来描述图像的特征。

通常 LBP 算子以每个像素点的 3×3 邻域为窗口，假设中心点的灰度值为 g_c ，邻域点的值为 g_0, g_1, \dots, g_7 ，则该区域内灰度值的联合分布可以表示为：

$$T = t(g_c, g_0, \dots, g_7) \quad (3-26)$$

在计算图像中每一个像素点的 LBP 值时，将该点的灰度值作为阈值， 3×3 邻域内另外 8 个点的灰度值分别与阈值比较，即：

$$T = t(g_c, g_0 - g_c, \dots, g_7 - g_c) \quad (3-27)$$

由于联合分布 T 的取值比较广泛，不利于描述，因此将灰度做差的结果用两个值来表示，即若中心阈值与其邻域像素点的灰度值之差值为负数，则将邻域中该点值记为 0，否则记为 1：

$$T \approx t(s(g_0 - g_c), s(g_1 - g_c), \dots, s(g_7 - g_c)) \quad (3-28)$$

其中 s 为符号函数，即 $s(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$ 。

给每个像素点乘以一个权值 2^i 并求和，就可以得到中心点周围区域纹理的描述值，称为 LBP 值，计算方法为：

$$LBP_8 = \sum_{i=0}^7 s(g_i - g_c)2^i \quad (3-29)$$

图 3-2 为 LBP 描述子的计算过程。

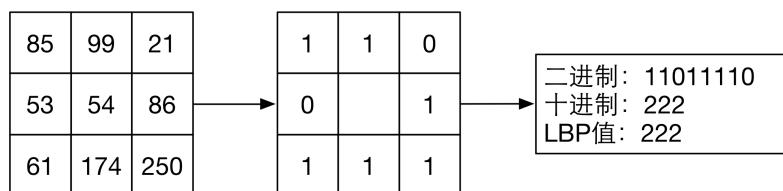


图 3-2 计算 LBP 值

经过上述过程后就能够得到图像中所有像素点的 LBP 值，然后用直方图统计图像中 LBP 值的出现频率，用其描述目标图像的纹理特征。综上采用 LBP 算法提取图像纹理特征的基本过程如下（如图 3-3）：

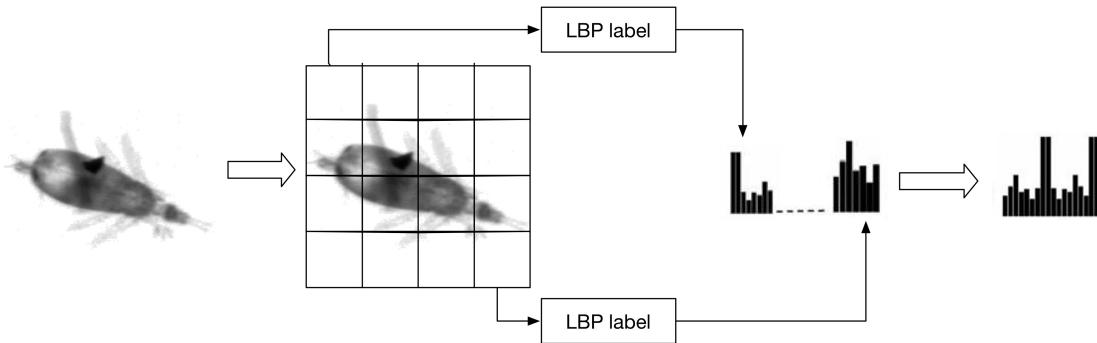


图 3-3 LBP 特征提取过程

1. 先将图像分为多个子区域。
2. 分别计算子区域中每个像素点的 LBP 值。
3. 用直方图分别统计每个子区域内 LBP 值的出现频率，对每个子区域内的纹理特征进行描述。
4. 将一幅图像中全部子区域的直方图串联到一起，就可以得到描述该幅图像的 LBP 特征向量。

3.1.3.4 二元梯度轮廓

二元梯度轮廓 (Binary Gradient Contours, BGC) 由 Fernandez 等人在 2011 年提出^[22]，是一种纹理特征描述算子。二元梯度轮廓与局部二值模式相似，它的基本思想是用 BGC 值对图像中每个像素与其 3×3 邻域内像素梯度的差别进行描述，通过直方图统计 BGC 值的出现频率来描述图像的纹理特征。

二元梯度轮廓与局部二值模式的不同之处在于，二元梯度轮廓沿一定封闭路径计算中心像素的 8 邻域像素之间的梯度，然后使用 0 作为阈值，若两像素间梯度值大于 0 则记为 1，否则记为 0。在这个过程中，计算梯度的路径有多种，其中典型的三种路径如图 3-4 所示。

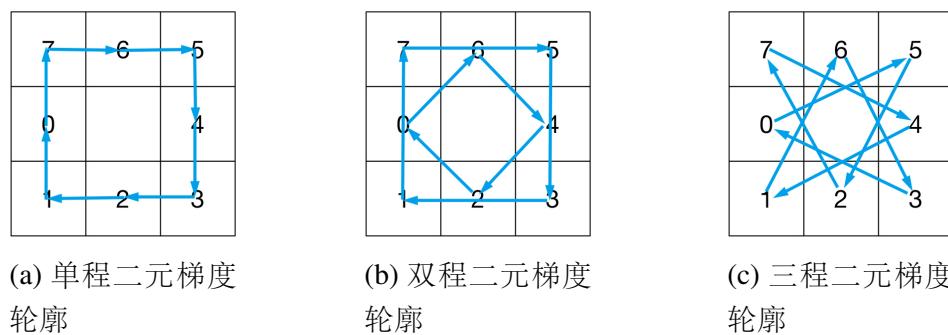


图 3-4 二元梯度轮廓的路径

图 3-4(a)由一条路径构成，叫做单程二元梯度轮廓 (BGC1)，公式如下：

$$g_1 = \begin{bmatrix} s(I_7 - I_0) \\ s(I_6 - I_7) \\ s(I_5 - I_6) \\ s(I_4 - I_5) \\ s(I_3 - I_4) \\ s(I_2 - I_3) \\ s(I_1 - I_2) \\ s(I_0 - I_1) \end{bmatrix} \quad (3-30)$$

图 3-4(b)中有两条路径，叫做双程二元梯度轮廓 (BGC2)，将两个回路计算的结果连接在一起共同来表示每一个点：

$$g_{2_1} = \begin{bmatrix} s(I_6 - I_0) \\ s(I_4 - I_6) \\ s(I_2 - I_4) \\ s(I_0 - I_2) \end{bmatrix}, \quad g_{2_2} = \begin{bmatrix} s(I_7 - I_1) \\ s(I_5 - I_7) \\ s(I_3 - I_5) \\ s(I_1 - I_3) \end{bmatrix} \quad (3-31)$$

$$g_2 = \begin{bmatrix} g_{2_1} \\ g_{2_2} \end{bmatrix} \quad (3-32)$$

同样，图 3-4(c)叫做三程二元梯度轮廓 (BGC3)，其表达式为：

$$g_3 = \begin{bmatrix} s(I_5 - I_0) \\ s(I_2 - I_5) \\ s(I_7 - I_2) \\ s(I_4 - I_7) \\ s(I_1 - I_4) \\ s(I_6 - I_1) \\ s(I_3 - I_6) \\ s(I_0 - I_3) \end{bmatrix} \quad (3-33)$$

因此每个像素都可以获得一个 8 位二进制二元梯度轮廓值，求取图像点的各

种二元梯度轮廓值的公式如下：

$$BGC_1 = w_8^T g_1 - 1 \quad (3-34)$$

$$BGC_2 = 15w_4^T g_{2_1} + w_4^T g_{2_2} - 16 \quad (3-35)$$

$$BGC_3 = w_8^T g_3 - 1 \quad (3-36)$$

其中， $w_j^T = [2^{j-1} 2^{j-2} \dots 2^1 2^0]$ 。

根据以上式子可以求出图像中每个像素点的二元梯度轮廓值，然后用直方图统计图像中二元梯度轮廓值的出现频率。

3.1.4 局部特征

3.1.4.1 内距离形状上下文

内距离形状上下文（Inner-Distance Shape Context, IDSC）由凌海滨在 2007 年提出^[23]，它是对形状上下文方法（Shape Context, SC）^[24]的一种改进。形状上下文是一种描述目标形状特征的方法，在 2002 年由 Serge Belongie 等人提出，该方法通过考察目标物体边缘上的点之间的空间位置关系来描述目标的形状特征，具体实现过程如下。

- 首先提取目标物体（如图 3-5）的轮廓边缘（如图 3-6）。由于轮廓上的像素点较多，需要从中采样 n 个点 $P = \{p_1, \dots, p_n\}$ 来近似表示目标物体的轮廓（如图 3-7）。在采集这 n 个点时要尽量保证采样点的重心和目标物体重心重合。



图 3-5 目标图像

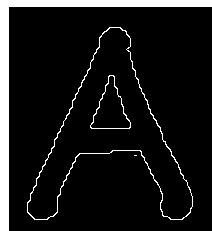


图 3-6 轮廓边缘



图 3-7 采样点

- 假设图 3-7 中第 i 个采样点为 p_i ，则连接这个采样点 p_i 与其他 $n-1$ 个点可以构成 $n-1$ 个向量，这些向量可以表示该点 p_i 与轮廓上其他点间的相对位置关系。以 p_i 为原点建立对数极坐标系，如图 3-8 所示，将图像中的坐标变换为极坐标，公式如下：

$$r = \sqrt{(x - x_0)^2 + (y - y_0)^2}, \quad \theta = \arctan\left(\frac{y - y_0}{x - x_0}\right) \quad (3-37)$$

其中 r 为欧式距离。将该极坐标系的半径 $\log r$ 和测量角度 θ 划分为 5、12bin。这个对数空间将被分为 48 个区域，用直方图 h_i 统计边缘上其他 $n - 1$ 个点落在对数空间每个区域内的数量，即：

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\} \quad (3-38)$$

其中， k 为直方图中区域数量，直方图 h_i 就是点 p_i 的形状上下文，它可以表示点 p_i 与其他 $n - 1$ 个点之间的空间位置关系。图 3-9 为图 3-8 中点 p_i 的形状上下文。

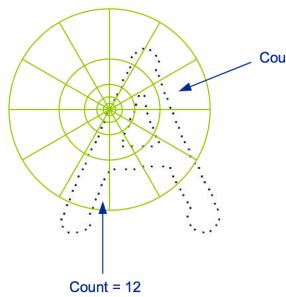


图 3-8 对数极坐标系

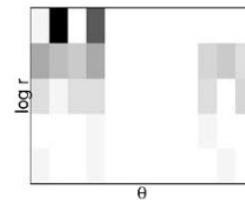


图 3-9 形状直方图

3. 计算表示目标边缘上所有采样点之间位置关系的形状上下文，得到的所有点的结果构成了对目标物体形状的描述。

在获得两个目标边缘上每一个采样点的形状上下文后，可以计算这两个目标上采样点之间的匹配关系，根据两目标上所有采样点之间的匹配程度可以得到两形状之间的相似程度，具体实现过程如下。

1. 假设形状 P 上的采样点为 p_i 和形状 Q 上的采样点为 q_j ，根据 $C_{ij} = C(p_i, q_j)$ 可以计算这两点的匹配代价。

$$C_{ij} \equiv C(p_i, q_j) = \frac{1}{2} \sum_{k=1}^n \frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)} \quad (3-39)$$

其中 C_{ij} 为点 p_i 和 q_j 匹配的代价值， $h_i(k)$ 和 $h_j(k)$ 表示点 p_i 和 q_j 的形状上下文。若 C_{ij} 值越小，则点 p_i 和 q_j 的形状上下文越相似，这两点的匹配程度越高^[25]。

2. 对两个形状轮廓边缘上的采样点进行匹配时，必须将采样点一一对应，并且

使全部点的匹配代价值的和最小，即

$$H(\pi) = \sum_i C(p_i, q_{\pi(i)}) \quad (3-40)$$

其中 π 是一种置换，约束条件是实现采样点的一一对应匹配，使用匈牙利算法可以解决。

3. 使用薄板样条 (Thin Plate Spline, TPS) 模型表示弹性坐标转换。用两个独立的 TPS 函数来模拟坐标的变换，从第一个形状的任意位置映射到第二个形状。

$$T(x, y) = (f_x(x, y), f_y(x, y)) \quad (3-41)$$

4. 计算两个形状之间的距离：

$$\begin{aligned} D_{sc}(P, Q) = & \frac{1}{n} \sum_{p \in P} \arg \min_{q \in Q} C(p, T(q)) + \\ & \frac{1}{m} \sum_{q \in Q} \arg \min_{p \in P} \arg \min_{p' \in P} C(p, T(q)) \end{aligned} \quad (3-42)$$

其中 T 为形状 Q 上的点到 P 的 TPS 变换。

内距离形状上下文是在形状上下文的基础上进行改进，即在内距离形状上下文中统计采样点的位置关系时使用内距离代替了形状上下文中的欧式距离。内距离是指目标轮廓上的两点位于其形状内部距离的最短值，如图 3-10 所示。内距离形状上下文相对应形状上下文而言，对目标的非刚性变化更加敏感，描述局部特征具有更好的鲁棒性。

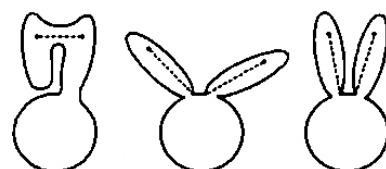


图 3-10 内距离

在本文研究中使用内距离形状上下文来提取浮游生物的形状特征，主要思想是通过内距离形状上下文计算待识别样本形状与已知浮游生物形状模板之间的相似度来描述形状特征，具体实现过程如下。

1. 针对每个数据集，选取 n 张不同形状的浮游生物图像，通过人工标注的方法得到这些图像中浮游生物的形状真值图作为形状模板。
2. 采用内距离形状上下文计算数据集中所有浮游生物图像和 n 张形状模板之间的距离，每张图像都会得到 n 个距离，这些距离表示每张图像的形状与形状模板之间的相似程度。针对每幅图像计算得到的 n 个值可以构成一个 n 维向量，这个向量可以描述该目标形状特征。

3.1.4.2 方向梯度直方图

方向梯度直方图（Histogram of Oriented Gradients, HOG）由 Dalal 等人在 2005 年提出^[26]，是一种特征描述方法，具有良好的旋转和平移一致性，被广泛的用于目标检测中。方向梯度直方图的基本思想是通过描述局部区域内的灰度梯度来表示图像的特征，其算法过程如下：

1. 先将图像进行灰度化，然后采用 Gamma 校正和灰度归一化处理获得的灰度图像。由于采集的图像受到光照变化、阴影的影响，通过 Gamma 校正来调节对比度，减少光照等其他因素的影响。

Gamma 压缩公式为：

$$I(x, y) = I(x, y)^{\text{gamma}} \quad (3-43)$$

2. 采用梯度算法计算所获得的灰度图像的梯度，获得目标的轮廓信息。

在计算图像梯度时，通常采用一阶微分模板 $[-1, 0, 1]$ 求梯度，对图像上任意一点用该模板可以得到垂直和水平方向上的梯度：

$$G_x(x, y) = H(x + 1, y) - H(x - 1, y) \quad (3-44)$$

$$G_y(x, y) = H(x, y + 1) - H(x, y - 1) \quad (3-45)$$

梯度的幅值和方向计算公式为：

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \quad (3-46)$$

$$\theta(x, y) = \tan^{-1} \frac{G_y(x, y)}{G_x(x, y)} \quad (3-47)$$

3. 将图像分为 Cell (即单元)，每个单元都由若干个像素组成。

4. 用直方图对图像中所有单元中的梯度方向加权统计。梯度的范围为 0 至 360 度，将其分为 n 份，统计梯度方向落在每份区域内的梯度加权值。
5. 每几个单元组成一个块（Block），将重叠单元块进行标准化，然后把块中所有单元的直方图向量组合起来得到的向量就是该块的特征描述符。
6. 图像中所有块的特征向量组合到一起就得到了该幅图像的 HOG 特征。

在本文实验中，使用方向梯度直方图来提取浮游生物的形态特征，在实验中将图像处理为 256×256 ，将单元大小设定为 32×32 。

3.1.4.3 尺度不变特征变换

尺度不变特征变换（Scale Invariant feature transform, SIFT）是一种局部特征描述算法，在 1999 年由 Lowe 提出^[27]，后来又进行了完善。SIFT 有旋转不变和尺度不变等性质，并且被广泛的用于目标识别、目标匹配等领域中。SIFT 算法的基本思想是提取图像中的高健壮性的特征点，通过对这些特征点的描述实现对图像中目标特征的描述，该算法的基本实现如下：

1. 采用高斯函数和图像下采样构建多尺度空间。

将原图像与不同参数的高斯函数卷积可以实现在不同尺度下对图像滤波：

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (3-48)$$

其中 $I(x, y)$ 表示原始图像， $L(x, y, \sigma)$ 表示滤波后得到的图像， $G(x, y, \sigma)$ 表示高斯函数，

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3-49)$$

上式中 x, y 是空间坐标，参数 σ 决定着对图像滤波程度。为了有效的确定图像中的关键点，将图像与两个相邻尺度高斯函数的卷积结果做差，从而构建高斯差分尺度空间（DOG）：

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (3-50)$$

采用以上方法来建立图像金字塔。图 3-11 为图像金字塔的结构，从中可以看出每个金字塔包括 i 个子八度（Octave），并且每个子八度有包括 s 层。子八

度是通过对图像下采样得到 i 个不同大小的图像（即不同尺度的图像，每次采样得到的图像都为原图的四分之一）组成的。每个子八度中的 s （通常为 3 至 5）层是使用高斯函数对图像进行不同程度滤波得到的。

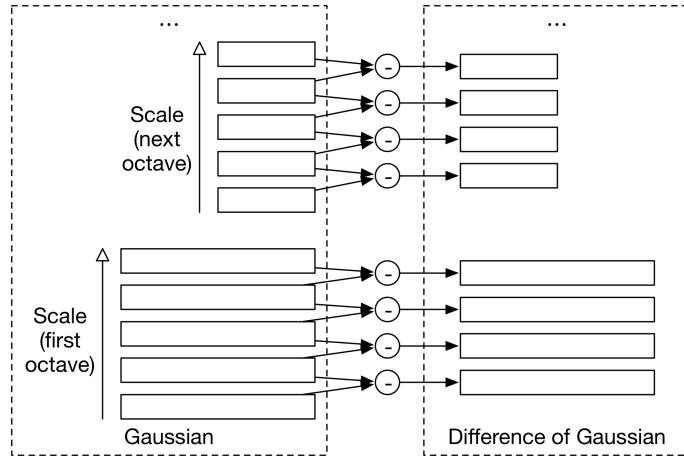


图 3-11 图像金字塔

2. 确定图像中的极值点。在 DOG 空间中确定极值点时，将每一个点与其空间中的邻域比较，观察该点是否是其邻域中的最大值或最小值点。如图 3-12 所示，每一个点的邻域包括其所在层的 8 邻域以及上下两层的 9 邻域，所以在确定极值点时每个点要与其空间邻域中的 26 个相邻点比较。

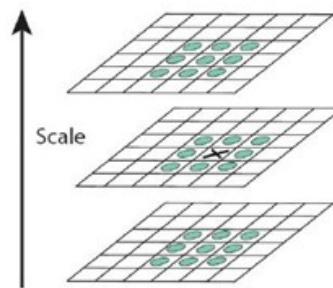


图 3-12 检测尺度空间的极值点

3. 精确定位特征点。得到的极值点中会存在着一些对比度较低和边缘响应点，将它们去除后保留下来的点就是特征点，即关键点。
在对关键点的位置和尺度进行精确定位时，采用拟合和三维二次函数将关键点中对比度较低的点和边缘相应点去除，从而增强特征描述的稳定性和抗噪能力，实现过程如下。

尺度空间的泰勒展开式如下：

$$D(x) = D + \frac{\partial D^T}{\partial x}x + \frac{1}{2}x^T \frac{\partial^2 D}{\partial x^2}x \quad (3-51)$$

对上式中 x 求导，且让导数为零可得到：

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x} \quad (3-52)$$

将上式代入 $D(x)$ 得：

$$D(\hat{x}) = D + \frac{1}{2} \frac{\partial D^T}{\partial x} \hat{x} \quad (3-53)$$

若 $|D(\hat{x})| \geq 0.03$ ，则该关键点保留，否则将其去掉。

另外还要排除图像中边缘上的关键点，位于图像横跨边缘处的关键点主曲率较大，而在竖直边缘处点的主曲率较小^[28]，通过这一性质来排除无用的关键点。主曲率可以通过 Hessian 矩阵求得：

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \quad (3-54)$$

若 H 的特征值为 α, β ，其中 α 较大， β 较小，则

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta \quad (3-55)$$

$$Det(H) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta \quad (3-56)$$

因此，令 $\alpha = \lambda\beta$ ，则：

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha + \beta)^2}{\alpha\beta} = \frac{(\gamma\beta + \beta)^2}{\gamma\beta^2} = \frac{(\gamma + 1)^2}{\gamma} \quad (3-57)$$

4. 确定关键点方向参数。在得到图像中关键点后，为了使其具有旋转不变性，需要确定它们的主方向。若关键点邻域内其他像素点 (x, y) 的梯度幅值和方向如下：

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (3-58)$$

$$\theta(x, y) = \tan 2 \frac{L(x, y + 1) - L(x, y - 1)}{L(x + 1, y) - L(x - 1, y)} \quad (3-59)$$

用直方图对关键点相邻区域内所有像素点的梯度方向进行统计，其横坐标的范围为 0 至 360 度。关键点的方向就是梯度直方图的峰值所表示的梯度方向。

5. 对关键点进行描述。在确定完关键点方向后，以关键点方向为轴建立坐标系生成关键点的特征描述子，具体过程如下：

- (a) 首先，将关键点作为中心点，取其 16×16 邻域，然后计算其中所有像素点的梯度。
- (b) 将这个邻域分为 16 个 4×4 的子区域，用 8 个 bin 的直方图加权统计每个子区域内像素的梯度方向。
- (c) 将得到的所有直方图结合在一起可以得到一个 128 (16×8) 维的向量，该向量就是对关键点的特征描述。

SIFT 算法通过提取并描述图像中的关键点来表示图像的特征。SIFT 对不同图像提取得到的关键点数量是不同的，每幅图可能包含成百上千个。因此，在进行目标识别过程中，采用 SIFT 特征来描述每幅图像时得到特征向量数量各不相同，特征点数量越多特征向量越多，分类计算量也越大。因此，用 SIFT 算法进行图像分类时通常要用到 Bag of Words 模型。

Bag of Words 模型也称为“词袋”，最初用于文本分类中，基本思想是假设两个文本，不考虑其中的语法、词序等，把文本看成一系列词汇的组合，将构成这两个文本的所有词汇放在一起，得到一个词袋，然后这两个文本又可以用这个词袋中的词汇重新描述。同理，将 SIFT 算法用于目标识别时，每幅图像可以看成一个文本，描述图像中每个特征点的特征向量看做为词汇，将所有图像中所有特征点的特征向量放在一起构成一个词袋，然后用这个词袋重新对每幅图像的特征进行描述。这样得到的描述每幅图像的特征向量不仅规整，而且减少了计算量，大大提高了计算效率。使用 BOW 模型描述图像特征的基本步骤如下：

1. 若数据集中有 n 张图像，采用 SIFT 算法提取所有图像中的特征点，并得到每一个特征点的 128 维特征向量。
2. 将 SIFT 提取的 n 张图像的所有特征点放在一起，采用 K-means 对这些特征点聚类，假设一共聚为 m 类。
3. 每个类别分别有一个聚类中心，计算每幅图像中所有特征点到这 m 个中心的距离，将每个特征点分别归到距其最近中心所属的类别中去，可以使用直方图向量统计每幅图像中特征点在 m 个类别中出现频率，该向量就是对图像特征的描述。

在本文研究中，采用 SIFT 提取数据集中全部图像的特征点，然后用 K-means 算法将所有特征点聚为 100 类，分别描述每幅图像中特征在 100 类中出现的频率，因此每幅图像的特征可以得到一个 100 维的特征向量。

3.2 浮游生物特征选择

特征选择是指去除特征中的冗余部分，降低特征维数，保留有用的特征，同时提高分类的效率和准确率。在分类识别中，有效的特征信息是训练优秀分类器的重要因素，特征中的冗余部分不仅会影响分类的结果^[29]，还会降低效率，因此在机器学习过程中对特征进行选择是至关重要的。

到目前为止，已经有很多研究者对特征选进行研究定义：在 1992 年 Kria 和 Rendell^[30] 提出特征选择是找到可以对目标进行识别的最小特征集合；后来 John 等人^[31] 提出减少特征维数是在能够提高或不要降低准确率的前提下进行；Koller 等人^[32] 认为在选择最小的特征集合时确保分类结果的分布与原始类分布相似；后来 Dash 等人^[33] 结合上述观点，将特征选择定义为在不降低分类准确度且不改变类比分布的情况下保留尽可能小的特征集合。

特征选择方法的基本步骤是：先生成候选的特征子集，然后对其进行评价，判断选择结果是否符合停止准则，若符合则对检验结果，否则重新生成候选特征子集。目前特征选择的方法有很多，按照搜索测量和评价标准的不同将其分类。

3.2.1 按搜索测量进行特征选择的方法

根据特征选择方法在选取子集时使用的搜索策略将其分为：全局最优、随机搜索和启发式搜索，下面对这三种方法分别进行介绍。

3.2.1.1 基于全局最优搜索策略的特征选择方法

分支界定是基于全局最优搜索策略的特征选择方法中可以获得最优结果的唯一算法^[29]，该算法的基本思想是：将所有可能的特征选择组合构成一个树状结构，按照特定的规则对树进行搜索，使搜索过程尽可能得到最优解而不必须遍历整个树。使用这种方法的前提是需要准则判据对特征有单调性，但是在处理高纬度特征时，该算法的时间复杂度较高。所以，基于全局最优搜索策略的算法虽然可以获得最优结果，但是很难被广泛的使用。

3.2.1.2 基于随机搜索策略的特征选择方法

基于随机搜索策略的特征选择方法通过有一定智能的随机搜索策略实现，在计算过程中该方法结合了特征选择与粒子群优化算法、模拟退火算法、遗传算法等，将采样和概率推理看做选择的基本，按照每个特征分类时的有效性，给它们分别赋予一个权值，然后按照设定或自适应阈值确定对分类有用的特征。若某个特征的权重超过该阈值，则这个特征便是有用的^[34]。

3.2.1.3 基于启发式搜索策略的特征选择方法

利用问题的启发信息作为引导，并进行搜索的方法称为启发式搜索，该方法可以降低问题的复杂度，并且减少搜索的范围。在特征选择过程中全局最优的搜索算法计算量可能很大，因此出现了以启发式搜索策略为基础的特征选择算法，该方法可以分为以下几种：单独最优特征组合、浮动搜索、增1去r选择方法、序列前向选择方法、广义序列前向选择方法、序列后向选择方法、广义序列后向选择方法、广义增1去r选择方法。虽然启发式搜索策略的效率较高，但是是以牺牲全局最优为代价。

3.2.2 按评价准则进行特征选择的方法

在特征选择过程中，对所选择特征好坏有不同的评价方法。根据评价方法是否依赖于后续的学习算法能够将特征选择方法分为两种：过滤式（Filter）和封装式（Wrapper）^[34]。过滤式特征选择独立于之后的学习算法，通常情况下可以直接使用训练数据的统计性能对特征进行评估，虽然计算效率较高，但是之后使用学习算法得到的结果和评估结果可能相差较大。封装式方法需要用后续的学习算法来对特征子集进行评价，因此结果与之后学习算法相差较小，但是计算量大。下面分别介绍这两种方法。

3.2.2.1 过滤式评价准则的特征选择方法

过滤式特征选择具有较高的效率，该方法通常用评价准则来增加特征与类别之间的相关性，去除掉不相关的杂质特征，优化特征子集，就像过滤器一样。根据评价准则可以将过滤式方法分成：一致性度量、依赖性度量、信息度量以及距离度量。这些方法的一个主要问题在找到的最优特征子集的规模往往较大，其中会包含一些噪声，但计算效率较高。

3.2.2.2 封装式评价准则的特征选择方法

封装式方法需要用之后的学习算法对选取的特征进行评估，将学习算法看做特征选择的一部分，按照学习得到的分类器性能进行特征选择。采用封装式方法进行特征选择时，用选取的特征集合来训练分类器，获得分类器的分类准确率可以作为评价所选特征重要性的标准。基于封装式的特征选择方法的计算速度要比基于过滤式的方法慢，但是它选择得到的特征子集维数较小，如今该方法在特征选择领域有较为广泛的应用。

在本文实验中，由于提取的图像特征种类丰富并且维数较高，因此其中会包含部分的冗余特征，这些特征不仅不利于分类准确率的提高，还降低了计算效率，因此在特征提取后又进一步进行了特征选择，采用的特征选择算法为基于封装式评价准则的特征选择算法。

3.3 本章小结

浮游生物种类繁多，类别相近的生物之间形态特征相差较小，而浮游植物和动物之间的差异又较大，这些特点增加了浮游生物分类的难度，因此需要全面充足的特征信息对浮游生物的形态特征进行描述。本章介绍了根据浮游生物形态特征选用的特征提取方法，包括：几何灰度统计特征；局部二值模式、二元梯度轮廓、Gabor 滤波器等纹理特征描述算法；内距离形状上下文、尺度不变特征变换等局部特征描述算法。这些特征提取算法可以对浮游生物的大小、形状、灰度、纹理等信息全面的描述。

由于提取的浮游生物特征较多，在提取的特征中可能存在冗余或不相关的信息，这些信息的存在不利于分类性能和准确率的提高，因此我们引入了基于封装式评价准则的特征选择方法对提取的特征进行筛选，去除冗余信息，降低维数，进而可以提高的分类器的泛化能力和整体性能。

4 基于多核学习的浮游生物图像分类研究

使用多核学习融合多种不同特征可以更好的发挥每种特征在分类过程中的积极作用。本章首先对多核学习的理论进行介绍；然后介绍设计的分类系统；最后，设计一系列对比实验分析分类系统的性能，并根据实验结果对分类系统性能进行评价。

4.1 多核学习理论

在图像分类识别过程中采用多种特征对目标进行描述有利于提高分类的准确率，而在这个过程中如何融合多种特征的方法又是人们关注的重点。通常特征融合算法可以分为决策级融合、特征级融合和数据级融合三类。数据级融合是将未加工的图像信息结合来得到更丰富的信息，常用的算法有主成分分析、线性加权法等。特征级融合是通过选择去除多余的特征，然后融合不相关的特征，例如聚类分析法、信息熵法等。决策级融合是将多个分类器结合，常用算法有贝叶斯融合、模糊聚类法、多核学习算法等。在本文中使用多核学习算法对提取的多种特征进行融合并训练分类器。

多核学习属于多视角学习。我们经常使用的支持向量机是一种单核学习方法，往往不能满足处理多种特征的需求，而多核学习为融合几种不同的核进行训练的学习方法，可以充分发挥每种特征在分类过程中的积极作用。下面以支持向量机为基础介绍核函数的原理，进而介绍多核学习的理论和方法。

4.1.1 支持向量机

早在 1995 年 Cortes 等人提出了支持向量机 (Support Vector Machine, SVM)^[35]，该方法属于监督学习，常用于目标分类和回归分析。支持向量机以 VC 维理论和结构风险最小化原理为基础，基本思想是在特征空间中构建一个最大间隔分类超平实现对未知样本的分类。因此，该方法适合处理二分类和回归问题，被普遍的应用在统计分类和回归分析中。

4.1.1.1 线性可分支持向量机

线性可分支持向量机是指训练样本在特征空间中线性可分，可以找到一个线性超平面将不同类别的样本完全正确的分开。设给定的样本为 $(x_1, y_1), \dots, (x_n, y_n)$ ，

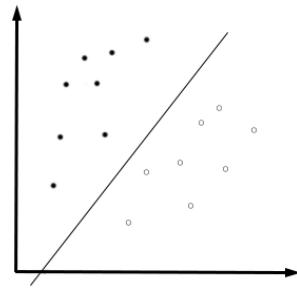


图 4-1 SVM 分类超平面

其中 $x_i \in X = R^n$, $y_i \in \{1, -1\}$, x_i 为第 i 个样本的特征向量, y_i 为第 i 个样本所属的类别。支持向量机对样本进行分类是通过在特征空间中找到最大间隔超平面将不同类别的样本划分开, 如图 4-1 所示, 超平面的方程可以表示为:

$$w^T x + b = 0 \quad (4-1)$$

其中, w 表示垂直于超平面的向量, b 表示超平面的截距, 该分类问题的决策函数为:

$$f(x) = s(w^T x + b) \quad (4-2)$$

若 $f(x_i)$ 大于 0, 则 x_i 为类别 1, 否则为类别 -1。

若要找到最大间隔超平面, 则需要计算样本到超平面的距离。通常 $|wx_i + b|$ 可以表示样本点 i 距离超平面的远近, 而 $wx_i + b$ 与 y_i 的符号相同, 因此可以用 $\hat{\gamma} = y_i(wx_i + b)$ 表示样本到分类平面的距离, 这就是函数间隔。然而在选择超平面时, w 和 b 等比例的变化并不会改变超平面, 但是却会改变函数间隔。因此, 引入几何间隔:

$$\tilde{\gamma} = \frac{\hat{\gamma}}{\|w\|} \quad (4-3)$$

一个超平面的几何间隔是所有训练样本距离该超平面的几何间隔中的最小值, 而支持向量机得学习目标是要在特征空间中找到一个超平面, 使其几何间隔最大。对线性可分的样本集来说, 其线性可分超平面有很多个, 然而几何间隔最大的分类超平面就一个。

若要找到最大间隔分类超平面, 需要令几何间隔最大, 即:

$$\max \tilde{\gamma} = \max \frac{\hat{\gamma}}{\|w\|} = \max \frac{y_i(w^T x + b)}{\|w\|} \quad (4-4)$$

其中函数间隔 $\hat{\gamma}$ 的大小并不会对上述最优化问题造成影响，因此令 $\hat{\gamma} = 1$ ，将其代入4-6，并且最大化 $\frac{1}{\|w\|}$ 和最小化 $\frac{1}{2}\|w\|^2$ 是等价的，得到求解支持向量机的最优化问题为：

$$\min \frac{1}{2}\|w\|^2, \quad s.t. \quad y_i(w^T x_i + b) \geq 1, i = 1, \dots, n \quad (4-5)$$

可以将上述最优化问题看做其原始问题，求解其对偶问题可以获得最优解。首先构建拉格朗日函数：

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^n \alpha_i [y_i(w^T x_i + b) - 1] \quad (4-6)$$

按照拉格朗日对偶性，原始问题可以转换为对偶问题的极大极小问题，即：

$$\min_{w, b} \theta(w) = \min_{w, b} \max_{\alpha_i \geq 0} L(w, b, \alpha) \quad (4-7)$$

求解以上对偶问题首先要固定 α 求 $L(w, b, \alpha)$ 关于 w, b 的极小，然后求对 α 的极大，就可以得到与之等价的对偶优化问题：

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle - \sum_{i=1}^n \alpha_i \quad (4-8)$$

$$s.t. \quad \sum_{i=1}^n \alpha_i y_i = 0 \quad (4-9)$$

$$\alpha_i \geq 0, i = 1, 2, \dots, n \quad (4-10)$$

在 KKT 条件成立的前提下，可以得到：

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad (4-11)$$

$$b = y_j - \sum_{i=1}^n \alpha_i y_i \langle x_i \cdot x_j \rangle \quad (4-12)$$

由此可以得到分类超平面和决策函数分别为：

$$\sum_{i=1}^n \alpha_i y_i <x \cdot x_i> + b = 0 \quad (4-13)$$

$$f(x) = s(\sum_{i=1}^n \alpha_i y_i <x_i, x> + b) \quad (4-14)$$

4.1.1.2 线性不可分支持向量机

对于分类中遇到线性不可分情况，需要把原始的特征映射到高维空间中，采用非线性变换将它转换为高维空间中的线性可分问题。假设低维空间的数据到高维空间的映射为 ϕ ，则式 4-6 可以转化为：

$$L(w, b, \alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j) \quad (4-15)$$

由于非线性映射比较复杂，因此引入核函数计算两个向量在映射到高维空间的内积函数：

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (4-16)$$

这样在对样本点进行预测时，只需要计算其在高维空间中的内积。引入核函数后，不必知道低维空间到高维空间映射 ϕ 的具体形式，能够根据原空间的函数计算得到内积。优化问题中引入核函数可得：

$$L(w, b, \alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (4-17)$$

同样可以得到决策函数为：

$$f(x) = s(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b) \quad (4-18)$$

4.1.2 核函数

核函数理论出现的较早，可以追溯到 1909 年的 Mercer 定理和之后再生核希尔伯特空间的出现。后来人们在研究势函数时将核函数应用到机器学习领域，之

后核函数在将线性支持向量机推广到非线性支持向量机时充分发挥了作用。

支持向量机在处理线性不可分的问题时，使用核函数将低维空间中的数据转换到高维空间中，通过在高维特征空间中建立超平面把样本区分开，从而实现将线性不可分问题转换成高维空间中的线性可分问题。在这个过程中，无需知道数据在高维空间中的具体形式，既避免“维数灾难”等问题的发生，也减少了计算的复杂度。

定义 4.1： 核函数：设 χ 是输入空间， H 为希尔伯特空间，如果存在一个从 χ 到 H 的映射 $\phi : \chi \rightarrow H$ ，使得对所有 $x_i, x_j \in \chi$ ，函数 $K(x_i, x_j)$ 满足条件

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

则称 $K(x_i, x_j)$ 为核函数。

使用核函数的优点是在训练分类器过程中，不需要知道 ϕ 的具体形式，可以直接使用核函数进行计算。目前比较常用的核函数有以下几种：

1. 线性核函数：

$$k(x_1, x_2) = \langle x_1, x_2 \rangle \quad (4-19)$$

2. 多项式核函数：是采用 n 阶多项式实现。

$$k(x_1, x_2) = (\gamma \langle x_1, x_2 \rangle + R)^n \quad (4-20)$$

3. 高斯径向基核函数：是最常用的核函数，可以把低维空间中的数据映射到无穷维。

$$k(x_1, x_2) = e^{-\gamma ||x_1 - x_2||^2} \quad (4-21)$$

4. Sigmoid 核函数：

$$k(x_1, x_2) = \tanh(\gamma \langle x_1, x_2 \rangle + R) \quad (4-22)$$

4.1.3 多核学习

4.1.3.1 多核学习原理

支持向量机是基于单核的分类算法，使用同一个核函数处理所有特征数据。然而不一样的核函数反应着的映射关系是不同的，因此针对不同数据使用不一样的核函数得到的分类效果会有一定差异。然而，在处理实际分类问题时，针对一幅图像提取一种特征往往是不够的，需要提取多种特征。显然，这时采用单个核函数并不能实现对所有特征的最优映射，在这种情况下，采用多个核函数比一个核函数更有助于提高分类的准确率。多核学习为每类特征指定核函数，给每个核赋予适合的权重，然后将所有的核函数组合到一起。在这个过程中，多核学习的研究重点是如何选择基本核及其权重。

构造多核学习模型的过程就是为每类特征选定合适的基本核，然后将它们组合到一起形成合成核。多核学习算法结构如图 4-2 所示。从图中可以看出，对于输入的数据中的不同特征分别为其选定适合的核函数，计算每个核函数的对应权重，按照一定的规则结合到一起形成合成核。合成核的基本形式为：

$$K(x_i, x_j) = f(k_m(x_i^m, x_j^m)_{m=1}^P) \quad (4-23)$$

其中 $k_m(x_i, x_j)$ 是基本核函数； m 表示使用的基本核数量； f 表示核结合函数，它决定基本核函数的结合形式，可以是线性或非线性函数。

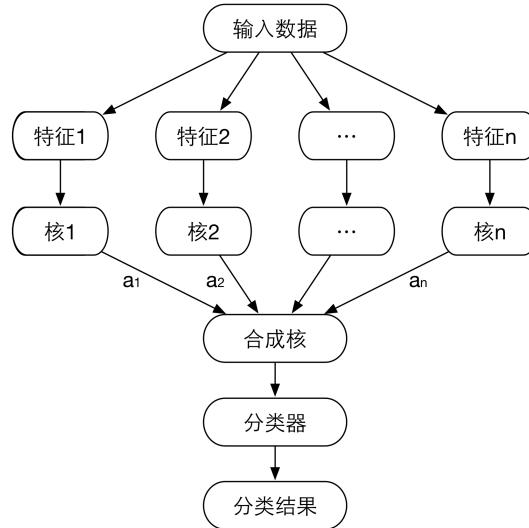


图 4-2 多核学习算法结构

多核学习有六个关键属性：（1）确定结合函数的方法，（2）结合核函数的形

式, (3) 确定结合函数参数的目标函数形式, (4) 计算结合函数参数的训练方法, (5) 基本学习算法, (6) 多核学习的计算复杂度。根据上述不同属性可以将多核学习算法进行分类^[36]:

1. 现有的多核学习算法用不同的学习方法来确定核结合函数的形式, 根据方法的不同可以将这些算法分为 5 类: 固定形式结合方法、启发式方法、最优化方法、贝叶斯方法和 Boosting 方法。
 - 固定形式是指结合函数中没有任何参数, 不需要训练, 例如将核以固定的形式相加或相乘。
 - 启发式方法采用有参数的结合函数组合基本核, 根据核矩阵或每个核单独使用时的表现来确定函数的参数。
 - 最优化方法也是使用有参数的结合函数, 通过解最优化问题来确定参数。
 - 贝叶斯方法将核结合函数的参数看做随机值, 先设定这些参数先验值, 然后通过推理学习参数。
 - Boosting 方法的思想来源于集成学习, 通过迭代增加新的核直到分类性能不再提高为止。
2. 多核学习算法根据核函数融合方式不同可以将其分为: 线性结合、非线性结合和数据依赖结合。
 - 线性结合是目前应用较广的方法, 主要包括无权重求和和有权重求和两种。
 - 非线性结合方法是指使用非线性函数做为结合函数, 例如包含相乘、幂运算的函数。
 - 数据依赖结合是指定特定的核权重给每组数据, 学习每个区域中最优核结合方式。
3. 确定结合函数参数能够根据优化不同目标函数得到, 现有的目标函数一般分为三类: 相似性函数、结构风险函数和贝叶斯函数。
 - 相似性目标函数是通过最大化相似度来确定结合函数的参数(相似度是根据训练集计算出结合核矩阵和最优核矩阵之间的相似性)。
 - 结构风险函数根据最小化正则项和误差项之和来确定结合函数的参数, 这称之为结构风险最小化。
 - 贝叶斯函数是用贝叶斯公式计算结合核的参数。
4. 计算结合函数参数的训练方法包括两种。一种是直接计算出结合函数参数和基础核函数参数; 另一种是通过迭代的方法实现, 先固定基础核函数的参数,

更新结合函数的参数，然后再固定结合函数的参数，更新基础核函数的参数，直到最终收敛就完成了训练。

5. 多核学习的基础学习算法有很多，例如支持向量机、核 Fisher 判别分析、核岭回归等。
6. 多核学习的计算复杂度主要依赖于训练方法和基础学习方法的复杂度。

同支持向量机这样的单核学习算法相比，多核学习在训练中不仅要计算 w, b 的值，还要得到每个核函数的权重。在这里最重要的问题是如何确定权重，近年来许多多核学习算法都是针对这一问题提出的。

多核学习算法中最经典的方法是简单多核学习 (SimpleMKL)^[37]，它被看做为多核学习算法的具体实现。后来人们还提出了广义多核学习 (Generalized Multiple Kernel Learning, GMKL)^[38]、局部多核学习 (Localized Multiple Kernel Learning, LMKL)^[39]、非线性多核学习 (Non-Linear Multiple Kernel Learning, NLMKL)^[40] 等算法，下面对简单多核学习和本文中使用的非线性多核学习算法进行介绍。

4.1.3.2 简单多核学习

SimpleMKL^[37] 使用线性结合的方式将多个基础核函数组合得到一个新的核函数，并采用梯度下降法求解，其核函数形式为：

$$K(x_i, x_j) = \sum_{m=1}^M d_m k_m(x_i, x_j), \quad \text{with } d_m \geq 0, \quad \sum_{m=1}^M d_m = 1 \quad (4-24)$$

上式中， k_m 表示基础核函数， M 表示基础核函数的数量， d_m 是基础核函数对应的权重系数，它表示着特征在分类过程中的重要性，是多核学习过程中需要求解的重要问题。根据支持向量机的原理，SimpleMKL 的原始优化问题可以等价为如下凸优化问题^[41]：

$$\begin{cases} \min \frac{1}{2} \sum_m \frac{1}{d_m} \|f_m\|^2 + C \sum_i \xi_i \\ s.t \quad y_i \sum_{m=1}^M f_m(x_i) + y_i b \geq 1 - \xi \quad \forall i \\ \xi_i \geq 0 \quad \forall i \\ \sum_{m=1}^M d_m = 1, \quad d_m \geq 0 \quad \forall m \end{cases} \quad (4-25)$$

上式中， f_m 为希尔伯特空间中的分类超平面。该式为多核学习的原始问题，要根据其确定参数 d_m 。采用拉格朗日算法可以将上述原始优化问题转换成对偶问题得到：

$$L = \frac{1}{2} \sum_m \frac{1}{d_m} \|f_m\|^2 + C \sum_i \xi_i + \sum_i \alpha_i (1 - \xi_i - y_i \sum_m f_m(x_i) - y_i b) \\ - \sum_i v_i \xi_i + \lambda (\sum_m d_m - 1) - \sum_m \eta_m d_m \quad (4-26)$$

其中 α_i 和 v_i 为拉格朗日乘子，而 λ 和 η_m 表示权重 d_m 的约束。接着分别对 L 自变量求偏导数，并令导数为 0 可得：

$$\begin{cases} \frac{1}{d_m} f_m(\bullet) = \sum_i \alpha_i y_i K_m(\bullet, x_i), & \forall m \\ \sum_i \alpha_i y_i = 0 \\ C - \alpha_i - v_i = 0, & \forall i \\ -\frac{1}{2} \frac{\|f_m\|^2}{d_m^2} + \lambda - \eta_m = 0, & \forall m \end{cases} \quad (4-27)$$

采用拉格朗日得到的对偶问题由于持续约束难以优化，这种约束可能转移到目标函数，但是目标函数变为不可微的，同样会造成求解困难。SimpleMKL 带约束的优化问题为：

$$\min_d J(d) \quad s.t. \quad \sum_{m=1}^M d_m = 1, d_M \geq 0 \quad (4-28)$$

$$J(d) = \begin{cases} \min_{f, b, \xi} \frac{1}{2} \sum_m \frac{1}{d_m} \|f_m\|^2 + C \sum_i \xi_i & \forall i \\ s.t. \quad y_i \sum_m f_m(x_i) + y_i b \geq 1 - \xi_i \\ \xi_i \geq 0 & \forall i \end{cases} \quad (4-29)$$

将上式可以转化为对偶问题：

$$J(d) = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_m m d_m K_m(x_i, x_j) + \sum_i \alpha_i \quad (4-30)$$

采用梯度下降法计算 $J(d)$ 的梯度下降表示，不断的更新 d 的值直到满足停止准则。根据支持向量机求解原理，SimpleMKL 的决策函数为：

$$f(x) = s \left(\sum_{i=1}^N \sum_{m=1}^M d_m \alpha_i y_i K_m(x_i, x_j) + b \right) \quad (4-31)$$

4.1.3.3 非线性多核学习

上面介绍的简单多核学习得到的分类器性能要比使用单个核的支持向量机好，该方法实质上可以看成支持向量机使用多个不同核函数的线性组合。Gonen 等人

在 2011 发表的文章^[36] 中对多种多核学习算法的性能进行了分析，他们在多个数据集上进行实验发现，采用非线性多核学习更有利于改善分类的性能。本文设计的浮游生物分类系统正是使用了非线性多核学习进行特征融合和训练分类器。Cortes 等人^[40] 提出了非线性多核学习（Non-linear Multiple Kernel learning, NLMKL），该方法基于核岭回归（Kernel ridge regression, KRR）采用多项式函数将核进行融合，他们提出的结合核的形式如下：

$$K_{\eta}(x_i, x_j) = \sum_{q \in Q} \eta_{q_1 q_2 \cdots q_m} k_1 \cdots k_m \quad (4-32)$$

其中 $Q = \{q : q \in Z_+^m, \sum_{l=1}^m q_l \leq d\}$ 。然而上式中需要学习的参数较多，为了减少学习的复杂度，将其化简为：

$$K_{\eta}(x_i, x_j) = \sum_{q \in R} \eta_1^{q_1} \eta_2^{q_2} \cdots \eta_m^{q_m} k_1 \cdots k_m \quad (4-33)$$

其中 $R = \{q : q \in Z_+^m, \sum_{l=1}^m q_l = d\}$ 。例如，当 $d = 2$ 时，核函数为：

$$K_{\eta}(x_i, x_j) = \sum_{l=1}^m \sum_{h=1}^m \eta_l \eta_h k_l k_h \quad (4-34)$$

在该非线性多核学习中，结合核函数的权重通过解最小-最大优化问题来求得。

在浮游生物的分类识别过程中为了提高分类的准确度通常要融合多种特征。在遇到多种特征进行分类的问题时，只采用单个核函数进行分类并不能充分发挥每种特征的作用。多核学习是通过为每种特征选择分别选择核函数并融合在一起构建分类器，这种方法可以有效的处理特征并解决单一核函数存在的不足。本文中我们使用非线性多核学习^[40] 来设计浮游生物分类系统，下面该分类系统进行详细介绍。

4.2 基于多核学习的浮游生物图像分类系统

我们研究浮游生物自动分类系统的出发点是扩大其适用范围，提高分类性能。因此在设计分类系统时采用多种特征提取方法对浮游生物形态特征进行描述，并结合了特征选择和多核学习方法来构建分类模型。本文提出的浮游生物自动分类系统由以下四个部分组成：图像预处理、特征提取、特征选择、多核学习，其算法结构流程图如图 4-3 所示。

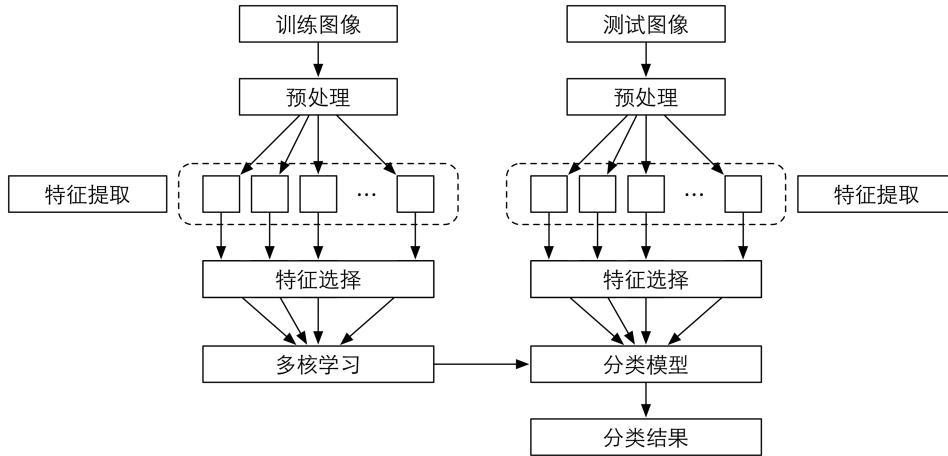


图 4-3 基于多核学习的浮游生物分类系统算法结构流程图

1、图像预处理

图像预处理是在图像识别之前进行的准备工作，包括去除图中的噪声、二值化分割图像背景等处理，避免无关因素对之后分类的影响。浮游生物数据集中的图像由水下图像采集设备获得，在采集浮游生物图像的过程中不可避免的会受到水中杂质等因素的干扰，造成采集的图像中会含悬浮物等噪声。因此为了提高图像的质量，在进行特征提取前要对采集的浮游生物图像进行预处理。

在本文实验使用的三个数据集中，WHOI 采集的浮游生物图像没有分割，因此在预处理时首先要对该数据集进行分割。对该数据集的图像通过检测边缘进行分割^[12]：首先将灰度图像进行相位一致性计算，然后采用 Canny 算子检测图像中的边缘，将获得的边缘图像用数学形态学算法（闭运算、膨胀、细化）进行处理，得到简单的轮廓边缘，使用获得的轮廓边缘就可以对其对应的浮游生物图像进行分割。

由于原浮游生物图像中可能存在悬浮颗粒等杂质，分割后图像中除了浮游生物外还会存在着小的杂质区域，因此在预处理过程中我们进行如下操作：首先得到原图像的二值图，然后采用数学形态学方法，去掉二值图像中面积小于一定像素数量的小连通区域；之后用得到的二值图对原图像重新做分割，获得去除噪声后的图像。图 4-4 中显示了预处理前后的浮游生物图像，能够看出图像 4-4(d)中的杂质点明显减少。

2、特征提取

采用 3.1 章节中提到特征描述方法对浮游生物的形态特征进行提取，可以实现从纹理、形状、灰度等不同角度对浮游生物形态特征进行描述。在特征提取后可以将提取的特征分为 10 类（每种特征提取方法提取的特征作为一类，其中粒子测度采用两组参数可以得到两组特征，将这两组特征分别作为一类），如图 4-5 所示。

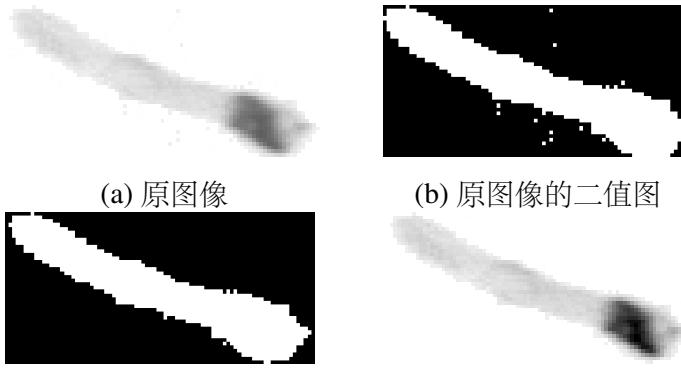


图 4-4 图像预处理

3、特征选择

在上一步中从不同角度提取了丰富的特征对浮游生物进行描述，然而这些特征中可能存在冗余或不相关的部分，它们不仅对分类性能的提高没有积极作用，甚至还会影分类效率。因此，使用特征选择从所有特征中选取最优子集、去除冗余部分，可以提高分类的性能和效率。并且，对不同数据集进行分类时，采用特征选择可以针对每一个数据集从所有特征中找到适合的特征组合。本文采用了基于封装式评价准则的特征选择方法^[42]，针对特征提取部分获得的 10 类特征分别进行选择，如图 4-5 所示。

4、多核学习

针对特征提取和特征选择后得到的 10 组特征信息，为每组特征分别选定 3 种核函数，采用多核学习的方法通过融合核函数实现特征的融合。在本文中使用非线性多核学^[40]将所有核函数融合到一起，同时训练得到最终分类器，如图 4-5 所示。最终得到的分类器可以对未知的浮游生物图像进行识别，并具有广泛的适用范围和较高的准确率。下面设计对比实验对分类系统的性能进行评价。

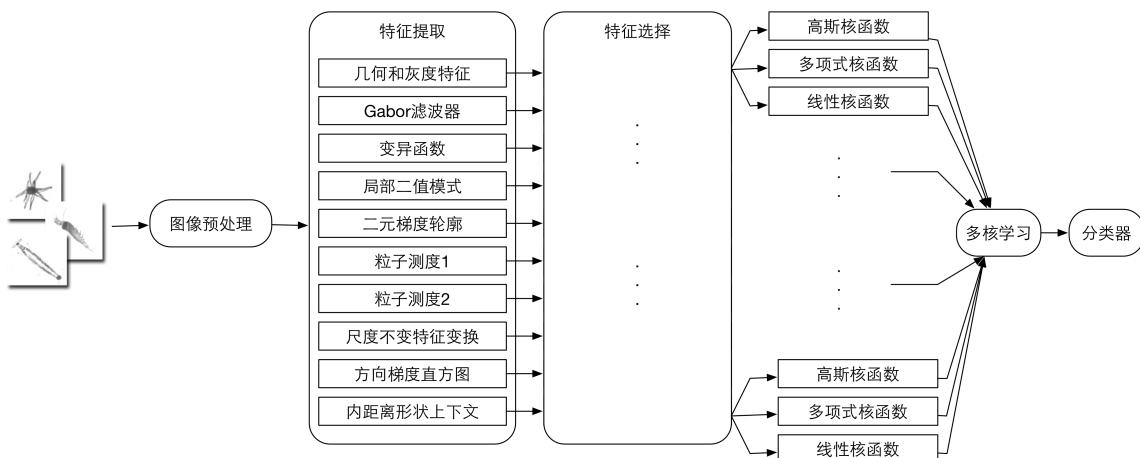


图 4-5 基于多核学习的浮游生物分类系统详细算法流程

4.3 对比实验

为了对本文设计的浮游生物分类系统的性能进行分析和评价，我们设计了如下三组对比实验：实验一是结合目前分类性能较好的浮游生物分类方法设计一个基准分类系统，作为分类系统性能对比评价的基准；实验二是在实验一的基础上，将使用 3.1 中提到的特征提取方法来对浮游生物特征进行描述；实验三采用本文设计的基于多核学习的浮游生物分类系统进行实验。将以上三个实验进行对比分析，从而实现对浮游生物分类系统各部分性能的评价。

4.3.1 基准实验

根据 Sosik 等人在 2007 年提出的浮游植物自动分类方法^[12] 和 ZooScan 系统^[13] 设计浮游生物分类的基准系统，该基准系统的算法流程框图如图 4-6 所示。

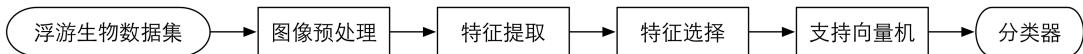


图 4-6 基准实验的算法流程框图

首先对浮游生物图像进行预处理，降低图像中的噪声信息。然后对浮游生物形态特征进行提取，这里使用的特征提取方法结合了 Sosik 论文中的 210 个特征和 ZooScan 系统中的 53 个特征，因此经过特征提取每幅图像可以获得一个 263 维的特征向量。在得到每幅图像的特征后，采用特征选择除去其中的冗余部分。经过特征选择后，三个数据集保留下来的特征维数如表 4-1 所示。最后对保留的特征归一化，用支持向量机算法来训练分类器，并使用混淆矩阵统计实验结果和分类准确率。

表 4-1 特征选择后得到的特征维数

	WHOI 数据集	ZooScan 数据集	Kaggle 数据集
特征维数	90	88	100

该实验在 2.2 中介绍的三个数据集上得到结果如表 4-2 所示，混淆矩阵如图 4-7（图中横轴和纵轴为数据集中浮游生物的种类，见附录 A；图中数值表示分为对应类别图像的数量，数值越大颜色越深）。在 WHOI 采集的数据集上分类结果的 F-Measure 为 0.8832，比相同条件下 Sosik 论文^[12] 中的分类结果有所提高。在 ZooScan 系统采集的数据集上得到分类结果的 F-Measure 为 0.8212，相比于 ZooScan 系统的分类结果（0.7947）^[13] 有一定提高。由此可以看出该实验设计的基准分类方法有较好的分类性能，可以作为评价浮游生物分类性能的基准。

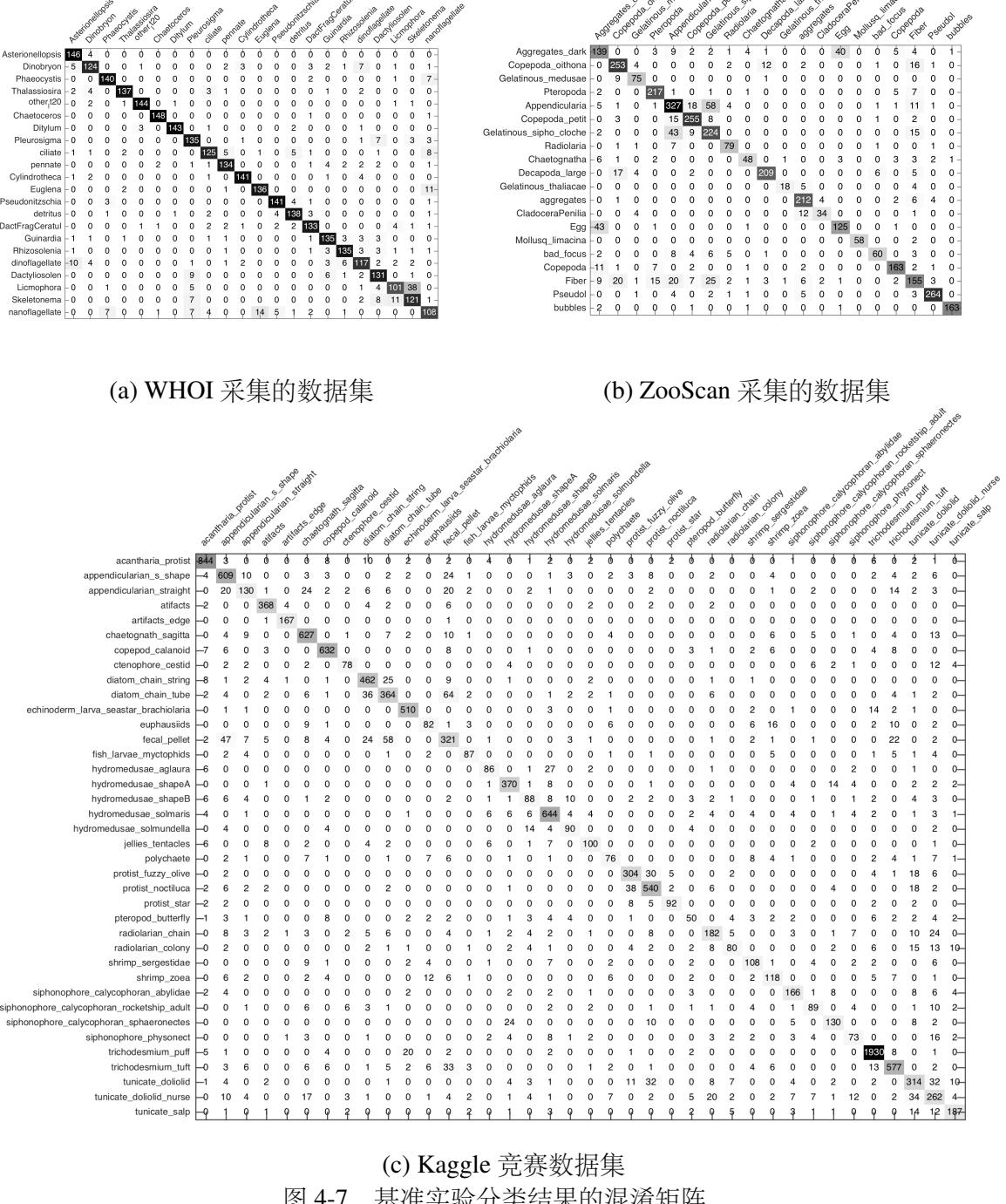


图 4-7 基准实验分类结果的混淆矩阵

4.3.2 特征对比实验

在基准系统的基础上设计特征对比实验，使用 3.1 中的特征提取方法来对浮游生物进行描述，从而对本文设计分类系统的特征提取部分的性能进行评价，该实验的基本算法流程如图 4-8。

首先进行图像预处理，然后采用 3.1 中的特征提取方法分别提取浮游生物的形状、纹理等特征。在完成特征提取后，将所有特征分为 10 类（每类特征提取方法

表 4-2 基准实验的分类结果

	WHOI 数据集	ZooScan 数据集	Kaggle 数据集
Recall	0.8827	0.8060	0.7536
Precision	0.8837	0.8370	0.7851
F-Measure	0.8832	0.8212	0.7690

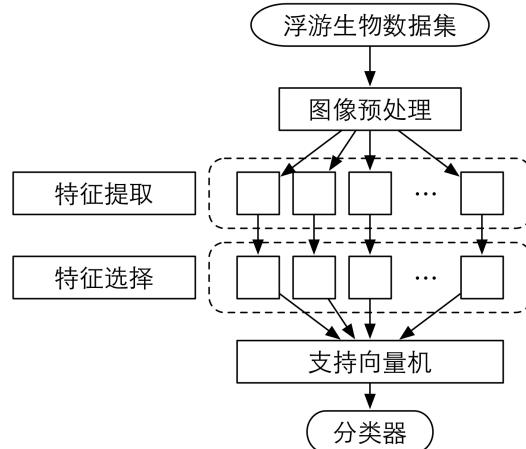


图 4-8 特征对比实验算法流程框图

提取得到的特征作为一类，其中粒子测度采用两组参数提取特征得到两组特征）。针对每类特征分别采用特征选择方法去除冗余特征，然后将经过特征选择后的到的所有特征串联在一起，用支持向量机训练得到最终的分类器。

在三个数据集上进行实验时，支持向量机采用不同参数得到的分类结果如表 4-3所示。从表 4-3中能够得到，本实验在三个数据集上得到的最优分类结果及其相应的参数为：当采用高斯核函数且惩罚因子 $C=100$ 时在 WHOI 取得最好分类结果，对应的 F 值为 0.8963；在 ZooScan 系统采集的数据集上，当使用高斯核函数且 $C=10$ 时取得最优分类结果， F 值为 0.8609；在 Kaggle 数据集上进行实验时采用高斯核函数且 $C=10$ 取得最优分类结果， F 值为 0.8304。本实验在这三个数据集上取得的最优分类结果的混淆矩阵如图 4-9所示（图中横轴和纵轴为数据集中浮游生物的种类，见附录 A；图中数值表示分为对应类别图像的数量，数值越大颜色越深）。将实验结果与实验一的结果进行对比可以对本文设计的浮游生物分类系统的特征提取部分的性能进行评价。

4.3.3 基于多核学习的浮游生物图像分类实验

将本文提出的基于多核学习的浮游生物分类系统分别在三个数据集上进行实验。在实验中为了对多核学习的性能进行分析，针对多核学习算法中的核函数种类设计了两组实验：(1) 在多核学习时，每种特征只设定一种核函数，其算法流程

表 4-3 特征对比实验的分类结果

数据集	C	高斯核函数			多项式核函数			线性核函数		
		Recall	Precision	F-Measure	Recall	Precision	F-Measure	Recall	Precision	F-Measure
WHOI 数据集	1	0.8400	0.8457	0.8428	0.8897	0.8905	0.8901	0.8645	0.8659	0.8652
	10	0.8894	0.8900	0.8897	0.8945	0.8953	0.8949	0.8812	0.8822	0.8817
	100	0.8957	0.8970	0.8963	0.8842	0.8854	0.8848	0.8633	0.8641	0.8637
ZooScan 数据集	1	0.7965	0.8394	0.8174	0.8245	0.8401	0.8322	0.7991	0.8186	0.8087
	10	0.8539	0.8680	0.8609	0.8414	0.8478	0.8446	0.8552	0.8399	0.8475
	100	0.8487	0.8638	0.8562	0.8304	0.8398	0.8351	0.8227	0.8177	0.8202
Kaggle 数据集	1	0.7726	0.8104	0.7910	0.7748	0.8040	0.7891	0.7132	0.7491	0.7307
	10	0.8241	0.8367	0.8304	0.8070	0.8192	0.8131	0.7844	0.7937	0.7890
	100	0.8209	0.8311	0.826	0.7901	0.8027	0.7964	0.7810	0.7795	0.7802

如图 4-10 所示，该组实验使用不同核函数和参数得到的分类结果如表 4-4 所示；(2) 每种特征使用三种核函数（分别为高斯核函数、多项式核函数、线性核函数），该实验的算法流程图如图 4-11 所示，不同参数得到的分类结果如表 4-5。

表 4-4 基于多核学习的浮游生物分类结果（一种核函数）

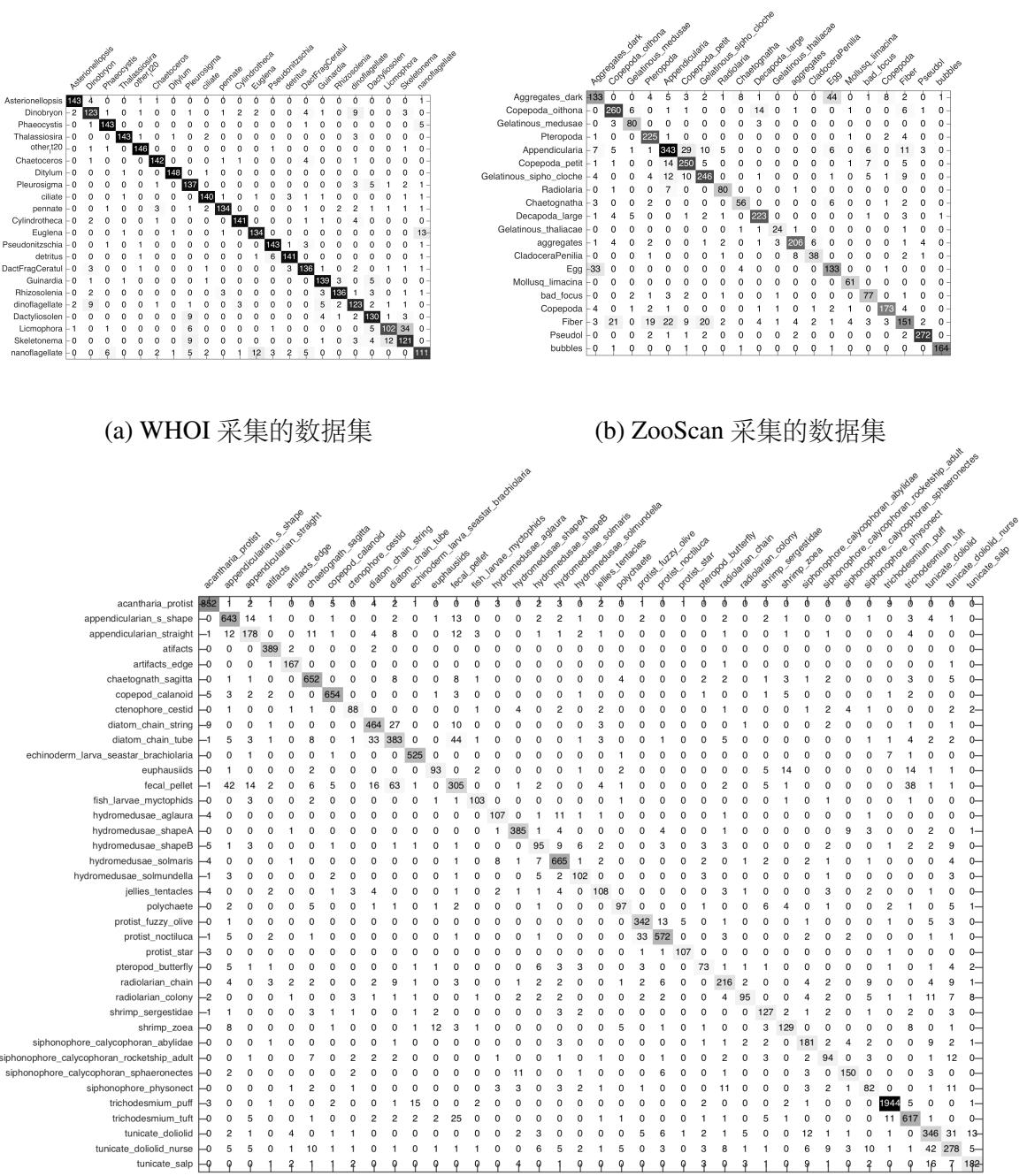
数据集	C	高斯核函数			多项式核函数			线性核函数		
		Recall	1-Precision	F-Measure	Recall	1-Precision	F-Measure	Recall	1-Precision	F-Measure
WHOI 数据集	1	0.8848	0.8865	0.8856	0.8958	0.8979	0.8968	0.8855	0.8882	0.8868
	10	0.8875	0.8896	0.8885	0.8967	0.8984	0.8975	0.8912	0.8935	0.8923
	100	0.8858	0.8880	0.8869	0.8939	0.8956	0.8947	0.8842	0.8859	0.8850
ZooScan 数据集	1	0.8332	0.8826	0.8572	0.8394	0.8739	0.8553	0.8178	0.8447	0.8310
	10	0.8626	0.8999	0.8809	0.8674	0.8824	0.8748	0.8398	0.8472	0.8435
	100	0.8660	0.9008	0.8831	0.8679	0.8837	0.8757	0.8351	0.8437	0.8394
Kaggle 数据集	1	0.7846	0.8259	0.8047	0.8039	0.8324	0.8179	0.7809	0.8076	0.7940
	10	0.8295	0.8358	0.8326	0.8260	0.8438	0.8348	0.8132	0.8234	0.8183
	100	0.8297	0.8316	0.8306	0.8211	0.8418	0.8313	0.7968	0.8090	0.8029

表 4-5 基于多核学习的浮游生物分类结果（三种核函数）

数据集	C	高斯核函数 + 多项式核函数 + 线性核函数		
		Recall	1-Precision	F-Measure
WHOI 数据集	1	0.8964	0.8983	0.8973
	10	0.8988	0.8997	0.8992
	100	0.9000	0.9009	0.9004
ZooScan 数据集	1	0.8542	0.8862	0.8699
	10	0.8834	0.9042	0.8937
	100	0.8831	0.9019	0.8924
Kaggle 数据集	1	0.8030	0.8388	0.8205
	10	0.8367	0.8551	0.8458
	100	0.8346	0.8512	0.8428

根据表 4-4 能够发现，采用一种核函数在三个数据集上取得的最优分类结果为：在 WHOI 数据集上进行实验，采用多项式核函数且 $C = 10$ 时得到的最优结果，F 值为 0.8975；在 ZooScan 采集的数据集上，采用高斯核函数且 $C=100$ 时取得最优分类结果，F 值为 0.8831；在 kaggle 数据集上取得最优分类结果的 F 值为 0.8348，对应的参数为多项式核函数且 $C=10$ 。图 4-12 为以上最优分类结果对应的混淆矩阵

基于多核学习的浮游生物图像分类研究



(图中横轴和纵轴为数据集中浮游生物的种类, 见附录 A; 图中数值表示分为对应类别图像的数量, 数值越大颜色越深)。分析表 4-5可以得到使用三种核函数在三个数据集上的最优分类结果: WHOI 数据集 F 值为 0.9004; ZooScan 数据集 F 值为 0.8937; Kaggle 数据集 F 值为 0.8458。将本实验与之前的实验结果进行对比可以对基于多核学习的浮游生物分类系统的整体性能进行评价。

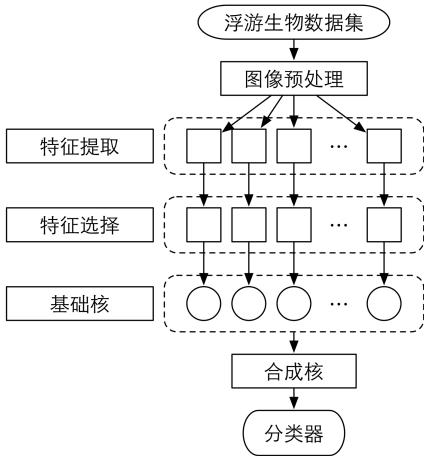


图 4-10 基于多核学习的浮游生物分类的算法流程图（一种核函数）

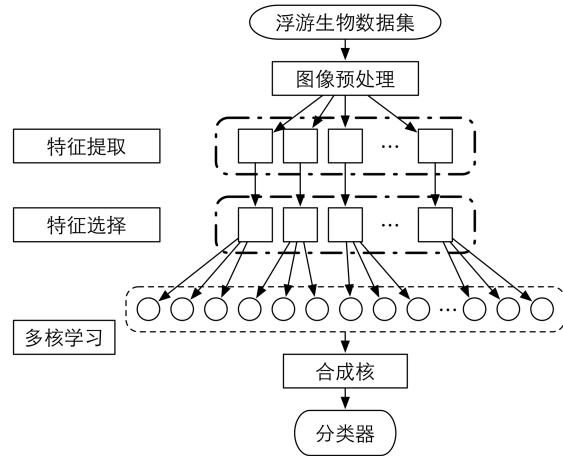


图 4-11 基于多核学习的浮游生物分类的算法流程图（三种核函数）

4.4 实验结果分析

浮游生物包括浮游植物和动物两大类，这两类生物之间的形态差别较大。同时，浮游生物的种类繁多，按照“界门纲目科属种”的级别进行划分，从“界”到“种”，越向下被归为同一属的生物之间形态相差越小。以上这两个方面是浮游生物分类研究两个较大的难点。为了扩大浮游生物分类系统的适用范围，提高分类的准确率，我们提出了一种基于多核学习的浮游生物分类系统，该系统具有以下特点：首先，从多个角度对浮游生物的特征进行描述，具有较广的适用范围；其次，采用特征选择算法去除冗余特征；最后使用多核学习训练分类器，提高分类器的性能。

为了对本文提出的系统性能进行评价，在4.3中设计了三组对比实验。实验一为基准实验，综合目前性能较好的浮游生物分类方法设计而来，作为评价分类系统对比的基准。实验二是在实验一的基础上使用3.1中的方法提取浮游生物特征，通过对比实验一与实验二的分类结果对特征提取方法的性能进行评价。实验三将本文设计的浮游生物分类系统在三个数据集上进行实验，将其分类结果与实验一和二对比来对分类系统总体性能进行评价。

根据表 4-2和表 4-3，比较实验一与实验二在三个数据集上取得的最优分类结果可得表 4-6。从表 4-6中能够得出，实验二在三个数据集上取得的分类准确率比实验一都有一定程度的提高，同时错误率也相应的降低，其中在 Kaggle 数据集上的效果最为明显， F 值提高了 0.0641，分类性能有明显提高。此外，对比图 4-7和图 4-9可以看出：与实验一相比，实验二得到的分类结果，在 Kaggle 竞赛数据集上有 36 个类别的准确提高了，在 WHOI 和 ZooScan 采集的数据集上分别有 14、16 个类别提高。由此可以看出，结合多种特征对浮游生物的形态特征进行全面描述有利浮游生物的区分。例如，图 4-7(b)中“Appendicularia”类图像中有 58 张图在实

	Asterionellopsis	Dinobryon	Phaeocystis	Thalassiosira	other120	Chaetoceros	Dilijum	Pleurosigma	pennate	Cylindrotheca	Euglena	Pseudoschizochloris	Diatom/Fragilaria	Guanidino	Rhizosolenia	Dinoflagellate	Lincostichus	Sphaerotilis	nanoflagellate
Asterionellopsis	141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Dinobryon	0	128	0	1	0	3	0	0	0	1	2	0	0	0	3	1	10	0	2
Phaeocystis	0	0	145	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Thalassiosira	0	0	0	142	0	1	1	0	3	1	0	0	0	0	0	2	0	0	0
other120	0	0	1	0	146	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Chaetoceros	0	0	0	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0
Dilijum	1	0	0	2	1	0	145	0	0	0	0	0	0	0	0	1	0	0	0
Pleurosigma	0	0	0	0	0	0	0	138	0	0	0	0	0	0	0	0	1	8	2
ciliate	0	0	0	1	0	0	0	0	146	1	0	0	0	0	1	1	0	0	0
pennate	0	1	2	0	4	0	1	1	134	0	0	0	0	0	2	3	1	1	0
Cylindrotheca	1	0	0	0	1	0	0	0	0	145	0	0	0	0	1	0	2	0	0
Euglena	0	0	1	0	0	1	0	0	0	134	0	0	0	0	1	0	0	0	13
Pseudoschizochloris	0	2	0	0	0	0	0	0	0	0	137	6	4	0	0	0	0	0	1
detritus	0	1	1	0	0	0	0	0	4	0	0	0	0	0	138	2	0	0	1
Diat/Fragilaria	0	1	0	0	1	2	0	0	2	0	0	0	0	0	138	1	0	0	2
Guanidino	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	138	5	0	0
Rhizosolenia	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	139	1	2
dinoflagellate	2	9	0	1	0	1	0	4	0	2	0	0	0	0	3	2	122	1	2
Dactylosolen	0	0	1	0	0	0	11	0	0	0	0	0	0	0	0	7	7	119	1
Licmophora	1	1	0	0	0	0	0	10	0	0	0	0	0	0	1	0	0	2	104
Skeletonema	0	0	0	0	0	0	13	0	0	0	0	0	0	0	2	0	1	2	15
nanoflagellate	0	0	11	1	0	2	4	1	0	0	12	0	2	3	0	0	0	1	113

(a) WHOI 采集的数据集

	acantharia_protist	appendicularian_s_shape	appendicularian_straight1	artifacts_edge	chaetognath_sagitta	copepod_calanoid	ctenophore_cestid	diatom_chain_cend	diatom_chain_string	euphausiids	fish_pellet	fish_larvae_myctophids	hydromedusae_aglaura	hydromedusae_shapeA	jellies_tentacles	polytropes	protofuzzy_olive	protist_star	pteropod_butterfly	radiolarian_chain	shrimp_colony	shrimp_sergeistidae	siphonophore_calyphoran_ablyidae	siphonophore_calyphoran_rockefeller_adult	siphonophore_calyphoran_sphaeronectes	siphonophore_physionect	trichodesmium_puff	trichodesmium_tuft	tunicate_doliolid	tunicate_doliolid_nurse	tunicate_salp			
acantharia_protist	860	3	0	3	0	3	2	0	2	0	20	1	0	1	2	4	1	0	1	2	0	0	1	0	1	2	2	1	0					
appendicularian_s_shape	-	634	13	1	0	2	2	0	2	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	5	0	1	0					
appendicularian_straight1	-1	14	184	0	9	0	1	1	7	0	0	12	2	0	0	1	0	0	0	0	1	1	0	0	0	0	5	0	1	0				
artifacts	-	0	0	383	3	0	0	4	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0				
artifacts_edge	-	0	0	0	2	168	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
chaetognath_sagitta	-	0	2	0	0	656	0	0	0	9	1	2	6	2	0	0	0	0	2	0	0	0	0	2	0	4	0	4	0					
copepod_calanoid	-4	4	1	4	0	0	650	0	0	0	2	4	0	0	1	0	1	0	0	0	0	1	0	0	1	3	0	0	0					
ctenophore_cestid	-	1	0	0	0	0	0	93	0	0	0	0	0	0	2	0	2	1	0	0	0	0	1	0	0	0	6	2	1	0	1			
diatom_chain_string	-5	0	2	0	2	0	0	1	461	27	0	0	15	0	1	0	0	0	3	0	0	0	2	0	0	0	0	0	0	1	0			
diatom_chain_tube	-2	2	5	2	0	8	0	0	42	365	0	0	53	1	0	0	1	0	0	2	0	1	0	0	0	0	0	0	0	1	0			
echinoderm_larva_seastar_brachio	-	0	2	0	0	0	1	0	0	0	525	0	0	0	0	0	1	0	0	0	0	0	0	0	0	5	7	2	5	0				
euphausiids	-	0	0	0	0	2	0	0	0	0	102	0	3	0	0	0	0	1	0	0	0	0	2	9	0	0	0	1	14	1	0			
fecal_pellet	-2	42	8	2	0	6	5	0	23	66	0	0	307	0	0	0	0	1	2	1	0	0	1	3	0	4	0	0	0	1	34	1	2	
fish_larvae_myctophids	-	0	1	1	0	0	2	0	0	0	1	0	1	104	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0		
hydromedusae_aglaura	-7	0	0	0	0	0	0	0	1	0	0	0	0	98	0	0	19	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0		
hydromedusae_shapeA	-	0	0	0	0	0	0	1	0	0	0	0	0	0	390	1	4	0	1	1	0	2	0	0	1	0	0	0	2	0	1	0		
hydromedusae_shapeB	-5	2	3	0	0	1	2	0	1	1	0	0	1	0	100	9	3	2	0	0	1	2	2	1	1	0	0	2	0	0	0	1	9	
hydromedusae_solimundella	-4	0	0	0	0	0	0	0	0	0	0	0	0	7	2	4	668	1	5	0	0	1	0	0	1	2	0	0	3	0	0	1	0	1
jellies_tentacles	-1	2	2	0	0	0	2	0	0	0	0	0	0	0	7	3	102	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	
polychaete	-	0	2	0	0	0	5	0	0	1	0	1	2	0	0	0	1	0	0	90	0	0	0	1	0	3	9	0	1	0	1	6		
protofuzzy_olive	-	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	333	21	4	0	0	0	0	1	0	0	0	2	6			
protist_noctilucia	-1	5	1	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	32	569	0	3	2	0	0	1	0	2	0	0			
protist_star	-5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	104	0	0	0	0	0	0	0	0	0	0			
pteropod_butterfly	-1	2	1	0	0	0	0	0	0	1	1	0	0	0	7	1	3	0	0	0	1	75	0	1	1	1	2	0	1	1	1	4		
radiolarian_chain	-6	0	3	1	2	0	0	3	6	0	0	4	0	0	1	3	0	1	1	6	0	0	226	2	0	0	1	2	1	4	0	6	6	
radiolarian_colony	-3	0	0	0	0	0	0	1	1	0	2	0	2	1	0	0	4	2	0	0	8	102	0	0	1	1	0	4	5	0	8	4		
shrimp_colony	-2	0	0	0	2	1	0	2	0	3	1	0	0	0	0	5	1	0	0	0	0	2	0	0	127	2	1	1	0	1	0	1		
shrimp_zoea	-5	0	0	0	0	0	0	0	0	15	3	1	0	0	0	0	0	7	0	0	0	0	0	3	133	0	0	0	1	5	0	1		
siphonophore_calyphoran_ablyidae	-	0	0	1	0	0	0	0	1	0	0	0	1	2	0	3	0	0	0	1	2	1	0	0	187	3	1	1	0	0	6	2		
siphonophore_calyphoran_rockefeller_adult	-	0	0	0	0	10	0	4	2	1	0	1	0	0	1	0	0	3	0	0	1	0	189	0	5	0	0	0	13	3				
siphonophore_calyphoran_sphaeronectes	-	0	0	0	0	0	0	1	0	0	0	0	0	0	12	0	1	0	0	0	10	0	0	0	0	3	0	4	0	0	3			
siphonophore_physionect	-	0	0	0	0	0	0	1	0	0	0	0	0	3	2	0	1	1	0	1	1	0	0	4	7	0	86	0	0	1	14	2		
trichodesmium_puff	-3	0	1	1	1	0	3	0	0	0	13	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	14	5	0			
trichodesmium_tuft	-1	1	4	0	0	1	2	0	1	3	2	2	29	0	1	0	0	0	2	0	0	0	1	0	2	5	0	0	0	13	607	0	1	
tunicate_doliolid	-2	2	0	0	0	1	0	1	0	0	0	0	0	0	3	1	0	0	0	5	3	1	1	6	7	0	0	5	0	1	0	0	347	
tunicate_doliolid_nurse	-4	5	0	0	0	9	1	1	0	5	0	0	4	1	0	1	5	2	3	1	2	1	0	5	11	2	2	1	9	5	1	11	1	4
tunicate_salp	-9	0	1	1	0	1	3	0	0	0	0	0	0	0	5	3	0	0	1	9	0	0	9	6	1	0	1	0	0	14	276	6		

(c) Kaggle 竞赛数据集

图 4-12 基于多核学习的浮游生物最优分类结果的混淆矩阵（一种核函数）

验一中被错误的分为“Gelatinous_sipho_cloche”类，而在实验二中，这个数字从 58 下降到了 10。因此，在分类过程中从多个角度的对浮游生物的形态特征进行描述，可以提高分类系统的性能。

实验三包含了两组实验：一个是在多核学习过程中每种特征只选用一种核函数；另一个是在多核学习时每种特征使用三种核函数。对比表 4-4 和表 4-3 中最优分类准确率可以的表 4-7，这是实验三中每种特征使用一种核函数的多核学习与实

表 4-6 特征对比实验与基准实验分类结果的对比

数据集	F-Measure 提高量
WHOI 数据集	0.0131
ZooScan 数据集	0.0397
Kaggle 数据集	0.0641

验二的对比结果，可以得出使用多核学习能够提高分类器的性能。观察混淆矩阵图 4-9 和图 4-12 可以发现：采用多核学习得到的分类结果，在 WHOI 采集的数据集中有 11 类图像分类结果得到了提高，在 Kaggle 数据集上有 21 类，而在 ZooScan 数据集上所有类别的准确率都得到了提高。由此可见，采用多核学习的分类系统相较于采用单核学习方法的分类系统有较好的性能。同时，对比表 4-4 和 4-5 得到表 4-8，可以发现使用多核学习时每种特征采用三种不同的核函数会使分类准确率有进一步的提高。

表 4-7 基于多核学习的浮游生物分类实验（一种核函数）与特征对比实验分类结果对比

数据集	F-Measure 提高量
WHOI 数据集	0.0012
ZooScan 数据集	0.0122
Kaggle 数据集	0.0044

表 4-8 每种特征使用三种核函数和一种核函数得到的分类结果对比

数据集	F-Measure 提高量
WHOI 数据集	0.0029
ZooScan 数据集	0.0106
Kaggle 数据集	0.0110

根据表 4-5 可得，我们提出的基于多核学习的分类系统在三个数据集的上最优分类结果的 F 值可以达到 0.9004、0.8937、0.8458，与实验一基准系统的最优分类结果对比有很大提高，如表 4-9。其中在 Kaggle 数据集上的性能提高最明显，F 值提高了 0.0768。由此可以看出，本文提出的基于多核学习的浮游生物分类系统有较好的分类性能和较高的泛化能力，可以广泛的应用于浮游生物研究。

4.5 本章小结

本文提出了基于多核学习的浮游生物分类系统，该系统主要采用多核学习融合多种特征进行分类识别，提高了分类系统的泛化能力和分类性能。本章介绍了

表 4-9 基于多核学习的浮游生物分类实验（三种核函数）与基准实验分类结果对比

数据集	F-Measure 提高量
WHOI 数据集	0.0172
ZooScan 数据集	0.0725
Kaggle 数据集	0.0768

多核学习的原理，采用非线性多核学习来进行特征融合和训练分类器。为了检验分类系统的分类性能，我们还设计了三组实验进行对比。首先根据 Sosik 等人提出的浮游植物分类系统^[12] 和 ZooScan 系统^[13] 设计基准分类系统，以此作为后续实验对比的基准；其次还设计了特征对比实验，对分类系统中特征提取部分的性能进行评价；最后采用本文设计的浮游生物分类系统进实验，对该分类系统的性能进行评价分析。在设计对比实验的同时为了检验系统的泛化能力，将所有的实验都分别在三个不同的数据集上进行。

5 总结与展望

5.1 总结

浮游生物自动分类系统是浮游生物检测的重要部分，它通过图像处理和模式识别方法对采集的浮游生物图像进行自动分类来实现。浮游生物自动分类可以解决人工分类专业水平要求高、费时、费力等问题，提高分类的效率，对浮游生物监测具有重要的研究意义。在本文中，我们从提高浮游生物分类系统的准确率、适用范围等方面出发，以特征提取、多核学习等知识为基础研究基于多核学习的浮游生物分类，主要工作内容如下：

1. 构建不同的浮游生物数据集，提高分类系统的适用范围。为了验证本文设计的分类系统的泛化能力，我们从网上搜集了不同机构和图像采集设备采集的浮游生物图像，构建三个不同的数据集进行实验，即包含浮游植物数据集，也包含浮游动物数据集。
2. 分析浮游生物的形态特征，选取合适的特征提取方法进行特征描述。人们对浮游生物进行分类时会根据其形状、纹理、大小等特征综合考虑。因此本文在设计浮游生物分类系统时，以人的识别方式为基础，分别提取了浮游生物的以下几类特征：几何灰度特征、纹理特征（例如 Gabor 滤波器、局部二值模式等）、局部特征（例如形状上下文、尺度不变特征变换等）、粒子测度。从多个角度对浮游生物进行全面描述，提取丰富的浮游生物特征信息，为后续分类做准备。
3. 从提取的特征中为每个数据集选取最优的特征组合，除去其中的冗余部分。采集的丰富的浮游生物特征中会存在不相关或冗余部分，它们会影响分类器的性能。在特征提取后采用特征选择去除其中的冗余部分，减少特征的维数，提高分类的准确率和泛化能力。
4. 分析多核学习理论，选用适合的多核学习方法训练分类器。在本文的分类系统中采用非线性多核学习，针对每种特征选定适合的核函数参数，实现特征的融合并充分发挥每种特征的分类性能。
5. 设计对比实验，建立合理的评价体系，对本文提出的分类系统各部分的性能进行评价。首先根据目前性能较好的浮游生物分类方法设计一个基准分类系统，作为对比评价的基准；然后进行特征对比实验，对分类系统特征提取部分的性能进行评价；最后，用本文设计的浮游生物分类系统进行实验，与之前两个实验对比，对多核学习以及整个分类系统的性能进行评价。

5.2 展望

本文提出的基于多核学习的浮游生物分类系统已经取得了较高的分类准确率，并且用于不同的浮游生物数据集时都有较好的性能，但是对浮游生物图像分类还有几个方面要进行深入研究：

1. 在多核学习中，每个核函数会对应生成一个核矩阵，核矩阵的数量越多计算量越大，在一定程度上会影响训练时间。因此多核学习相对于单核学习算法的计算速度较慢。如何提高多核学习的计算速度，使该系统可以更好的用于浮游生物实时监测是接下来需要研究的内容之一。
2. 从本文方法在三个数据集上的实验结果可以发现，虽然本文方法对不均衡的数据集（即数据集中每个类别图像数量不等）的分类准确率有一定的提升，然而其分类准确度还低于均衡数据集，因此下一步可以针对数据集不均衡问题进行深入研究。

参考文献

- [1] 孙晓霞,孙松. 海洋浮游生物图像观测技术及其应用. 地球科学进展, 2014, 29(6):748–755.
- [2] Grosjean P, Picheral M, Warembourg C, et al. Enumeration, measurement, and identification of net zooplankton samples using the zooscan digital imaging system. ICES Journal of Marine Science: Journal du Conseil, 2004, 61(4):518–525.
- [3] 毕永坤. 基于 zooscan 技术的浮游动物体形参数和生物量关系研究 [D]. 中国科学院研究生院, 2011.
- [4] Davis C, Gallager S, Berman M, et al. The video plankton recorder (vpr): design and initial results. Arch. Hydrobiol. Beih, 1992, 36:67–81.
- [5] Davis C S, Hu Q, Gallager S M, et al. Real-time observation of taxa-specific plankton distributions: an optical sampling method. Marine Ecology Progress Series, 2004, 284:77–96.
- [6] Sieracki C K, Sieracki M E, Yentsch C S. An imaging-in-flow system for automated analysis of marine microplankton. Marine Ecology Progress Series, 1998, 168:285–296.
- [7] Samson S, Hopkins T, Remsen A, et al. A system for high-resolution zooplankton imaging. IEEE Journal of Oceanic Engineering, 2001, 26(4):671–676.
- [8] Olson R J, Sosik H M. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging flowcytobot. Limnol. Oceanogr. Methods, 2007, 5(6):195–203.
- [9] Tang X, Stewart W K, Vincent L, et al. Automatic plankton image recognition. Artificial Intelligence for Biology and Agriculture. Springer, 1998: 177–199.
- [10] Zhao F, Tang X, Lin F, et al. Binary plankton image classification using random subspace. IEEE International Conference on Image Processing 2005, volume 1. IEEE, 2005. I–357.
- [11] Tang X, Lin F, Samson S, et al. Binary plankton image classification. IEEE Journal of Oceanic Engineering, 2006, 31(3):728–735.
- [12] Sosik H M, Olson R J. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol. Oceanogr. Methods, 2007, 5(204):e216.
- [13] Gorsky G, Ohman M D, Picheral M, et al. Digital zooplankton image analysis using the zooscan integrated system. Journal of Plankton Research, 2010, 32(3):285–303.
- [14] Mosleh M A, Manssor H, Malek S, et al. A preliminary study on automated freshwater algae recognition and classification system. BMC bioinformatics, 2012, 13(17):1.
- [15] Jeffrey E, Hongyu L, Mark D O. Quantifying california current plankton samples with efficient machine learning techniques. Proceedings of OCEANS 2015 MTS/IEEE Washington, 2015. 1–9.
- [16] 王怀亮. 交叉验证在数据建模模型选择中的应用. 商业经济, 2011, (10):20–21.
- [17] Matheron G. Random sets and integral equation, 1978.
- [18] 曹健渭, 卢荣胜, 雷丽巧, 等. 基于散斑纹理变差函数的平磨表面粗糙度测量技术. 仪器仪表学报, 2010, (10):2302–2306.
- [19] 高梓瑞. Gabor 滤波器在纹理分析中的应用研究 [D]. 武汉: 武汉理工大学, 2012.

- [20] 王铌. 海洋浮游生物数字图像处理方法 [D]. 中国海洋大学, 2006.
- [21] Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. *Pattern Recognition*, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on, volume 1. IEEE, 1994. 582–585.
- [22] Fernández A, Álvarez M X, Bianconi F. Image classification with binary gradient contours. *Optics and Lasers in Engineering*, 2011, 49(9):1177–1184.
- [23] Ling H, Jacobs D W. Shape classification using the inner-distance. *IEEE transactions on pattern analysis and machine intelligence*, 2007, 29(2):286–299.
- [24] 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.
- [25] 杨小娜. 基于形状上下文的目标形状识别与匹配 [D]. 昆明理工大学, 2013.
- [26] Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1. IEEE, 2005. 886–893.
- [27] Lowe D G. Object recognition from local scale-invariant features. *Computer vision*, 1999. The proceedings of the seventh IEEE international conference on, volume 2. Ieee, 1999. 1150–1157.
- [28] 陈健斌. 图像特征提取及其相似度的研究和实现 [D]. 西安电子科技大学, 2012.
- [29] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述. *控制与决策*, 2012, 27(2):161–166.
- [30] Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm. *Tenth National Conference on Artificial Intelligence*, 1992. 129–134.
- [31] John G H, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. *Machine Learning Proceedings*, 1998. 121–129.
- [32] Koller D. Toward optimal feature selection. *Proc. 13th International Conference on Machine Learning*, 2000. 284–292.
- [33] Dash M, Liu H. Feature selection for classification. *Intelligent Data Analysis*, 1997, 1(1):131–156.
- [34] 宁永鹏. 高维小样本数据的特征选择研究及其稳定性分析 [D]. 厦门大学, 2014.
- [35] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, 20(3):273–297.
- [36] Nen M, Alpayd, Ethem N. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 2011, 12:2211–2268.
- [37] Rakotomamonjy A, Bach F R, Canu S, et al. Simplemk1. *Journal of Machine Learning Research*, 2008, 9(3):2491–2521.
- [38] Varma M, Babu B R. More generality in efficient multiple kernel learning. *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009. 1065–1072.
- [39] Gönen M, Alpaydin E. Localized multiple kernel learning. *Proceedings of the 25th international conference on Machine learning*. ACM, 2008. 352–359.
- [40] Cortes C, Mohri M, Rostamizadeh A. Learning non-linear combinations of kernels. *Advances in neural information processing systems*, 2009. 396–404.

- [41] 孙锐, 侯能干, 陈效华. 基于多特征和多核学习的行人检测方法的研究. 图学学报, 2014, 35(6):869–875.
- [42] Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1-2):273–324.

附录 A 浮游生物种类

实验构建的三个数据集中浮游生物的分类：

表 A-1 WHOI 采集的数据集

浮游生物种类	
Asterionellopsis	Euglena
Dinobryon	Pseudonitzschia
Phaeocystis	Detritus
Thalassiosira	DactFragCeratul
Other_20	Guinardia
Chaetoceros	Rhizosolenia
Ditylum	Dinoflagellate
Pleurosigma	Dactyliosolen
Ciliate	Licmophora
Pennate	Skeletonema
Cylindrotheca	Nanoflagellate

表 A-2 ZooScan 系统采集的数据集

浮游生物种类	
Aggregates_dark	Gelatinous_thaliaceae
Copepoda_oithona	Aggregates
Gelatinous_medusae	CladoceraPenilia
Pteropoda	Egg
Appendicularia	Mollusq_limacina
Copepoda_petit	Bad_focus
Gelatinous_sipho_cloche	Copepoda
Radiolaria	Fiber
Chaetognatha	Pseudol
Decapoda_large	Bunnles

表 A-3 Kaggle 竞赛数据集

浮游生物种类	
acantharia_protist	jellies_tentacles
appendicularian_s_shape	polychaete
appendicularian_straight	protist_fuzzy_olive
artifacts	protist_noctiluca
artifacts_edge	protist_star
chaetognath_sagitta	pteropod_butterfly
copepod_calanoid	radiolarian_chain
ctenophore_cestid	shrimp_sergestidae
diatom_chain_string	shrimp_zoea
diatom_chain_tube	siphonophore_calycophoran_abylidae
echinoderm_larva_seastar_brachiolaria	siphonophore_calycophoran_rocketship_adult
euphausiids	siphonophore_calycophoran_sphaeronetes
fecal_pellet	siphonophore_physonect
fish_larvae_myctophids	trichodesmium_puff
hydromedusae_aglaura	trichodesmium_puff
hydromedusae_shapeA	trichodesmium_tuft
hydromedusae_shapeB	tunicate_doliolid
hydromedusae_solmaris	tunicate_doliolid_nurse
hydromedusae_solmundella	tunicate_salp

致 谢

转眼间三年的研究生生活即将结束，在这三年中我得到了导师、同门、家人以及朋友等许多人的帮助，在研究生生涯即将结束时，我想向这些给予我帮助的人们表示诚挚的感谢。

首先，感谢导师姬光荣老师和郑海永老师对我的细心指导。在这三年的研究生生活中，郑老师教会我如何学习科研、如何工作思考、如何做人做事，他的严格要求让我养成好的习惯，并不断的提高自己。同时，我也十分感谢姬老师，在这三年里对我进行悉心指导，并给予我许多的关心和照顾。我研究生生活中最幸运的事就是可以同时有这两位老师的指导和关心，十分感谢两位老师为我所付出的一切。

感谢朱亚菲师姐、孙晓庆师姐以及邱欣欣师姐，她们在科研和生活中给予我很大的帮助。感谢同级的伙伴们，你们的帮助陪伴伴随我走过了这美好难忘的三年。感谢实验室的师弟师妹们，这三年的时光因为你们而美好。

感谢我的父母，你们对我的支持和无私付出，在我面对困难时给予我力量。在以后的日子里，我会更加努力，定不辜负你们的期望。

感谢国家自然科学基金项目“基于视觉注意结合生物形态特征的海洋浮游植物显微图像分析”（批准号：61301240）、国家自然科学基金项目“基于生物形态特征的中国海常见有害赤潮藻显微图像识别”（批准号：61271406）、中央高校基本科研业务费项目“海洋浮游动物原位探测与分析系统”（批准号：201562023）的资助。

最后，感谢所有关心和帮助过我的人，祝愿你们永远幸福快乐！

个人简历、在学期间发表的学术论文与研究成果

个人简历

1992年5月14日出生于山东省威海市。

2010年9月考入浙江师范大学数理与信息工程学院电子信息工程专业，2014年7月本科毕业并获得学士学位。

2014年9月考入中国海洋大学电子系攻读硕士学位至今。

发表的学术论文

- [1] Wang R, Dai J, Zheng H, Ji G, Qiao X. Multi features combination for automated zooplankton classification, in Proceedings of OCEANS'16 MTS/IEEE Shanghai, 2016.
- [2] Gu Z, Wang R, Dai J, Zheng H, Zheng B. Automatic searching of fish from underwater images via shape matching, in Proceedings of OCEANS'16 MTS/IEEE Shanghai. 2016.

在学期间参加的研究项目

1. 国家自然科学基金项目“基于视觉注意结合生物形态特征的海洋浮游植物显微图像分析”(批准号：61301240)。
2. 国家自然科学基金项目“基于生物形态特征的中国海常见有害赤潮藻显微图像识别”(批准号：61271406)。
3. 中央高校基本科研业务费“海洋浮游动物原位探测与分析系统”(批准号：201562023)。