中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于深度学习的高清航拍图像目标检测

作者姓名：朱海港

指导教师：韩振军　副教授　中国科学院大学

学位类别：工学硕士

学科专业：计算机应用技术

研究所　：中国科学院大学电子电气与通信工程学院

2015 年 5 月

Research of High Resolution Aerial Object Detection Based on Deep Learning

By

Haigang Zhu

A Thesis Submitted to

University of Chinese Academy of Sciences

In Partial Fulfillment of the Requirement

For the Degree of

Master of Computer Application Technology

College of Electronic, Electrical and Communication

Engineering

05, 2015

中国科学院大学直属院系

研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

中国科学院大学直属院系

学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密的学位论文在解密后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘　要

航拍目标检测是利用目标检测算法检测航拍图像中的特定目标的技术，是计算机视觉领域的重要问题之一。航拍目标检测在军事目标智能识别，遥感影像解析以及民用航空等领域具有广阔的应用前景。航拍图像中的目标有以下特点：（一）航拍图像中的目标形态受颜色变化，长宽比变化以及复杂背景的影响较大；（二）航拍图像中目标的角度变化比较大，常规的手工设计的特征难以描述角度变化大的航拍目标。本文提出使用深度卷积网络的特征组合提取航拍目标的角度不敏感特征。本文研究了卷积网络的特征和目标角度之间的关系，并提出采用基于分割的感兴趣区域提取算法以改进航拍目标检测算法的框架。我们标定了 SDL 高清航拍数据集，并对数据集中的飞机和汽车进行了检测，验证了本文算法的有效性。

本文的主要贡献如下：

1. 本文提出了应用高维数据可视化工具来描述目标的特征与角度的关系。从卷积神经网络的特征分析着手，发现了卷积网络某些层的特征可作为角度不敏感特征描述子。

2. 基于角度不敏感特征描述子，本文提出了航拍图像的非旋转检测框架。该框架使用基于图像分割的算法进行感兴趣区域提取（粗定位），然后进行精确检测。这个框架比原本的旋转原图的框架在准确率上有了较大的改进。

3. 针对目标检测需要标定大量训练样本的问题，本文提出采用弱监督学习的方式进行正例挖掘，以减轻传统的手工标定带来的人力负担；并在航拍飞机数据集上做了正例挖掘实验，实验表明，本文的算法能够较好地挖掘出正例样本。

关键词： 航拍目标，目标检测，深度学习，弱监督学习

Abstract

Object detection in aerial images is one of the fundamental problems of the field of object detection, which is a significant problem of computer vision. An aerial object detection system can be widely used in reconnaissance and remote sensing image resolution as well as others. However, detecting objects in aerial images is challenged by variance of object colors, aspect ratios, cluttered backgrounds, and in particular, undetermined orientations. In this thesis, we propose to use Deep Convolutional Neural Network (DCNN) features from combined layers to perform orientation robust aerial object detection. We explore the inherent characteristics of DCNN as well as relate the extracted features to the principle of disentangling feature learning. An image segmentation based approach is used to localize ROIs of various aspect ratios, and ROIs are further classified into positives or negatives using an SVM classifier trained on DCNN features. With experiments on two datasets collected from Google Earth, this thesis demonstrates that the proposed aerial object detection approach is simple but effective.

Our contributions are as follows:

1. We propose to use a tool to visualize the relationship between feature and orientation. With this tool, we find that the feature extracted from some layer of convolution neural network is of orientation invariability.
2. Based on the feature, we proposed to use the feature to train a detector of orientation invariability instead of detection objects in rotated images. We propose to use segmentation based algorithm to extract regions of interest, and feed the feature of these regions into SVM. This pipeline is simple but effective.
3. We propose to use weakly supervised learning algorithm to select positive training samples instead of using human annotated samples. Our experiments show that our algorithm performs well.

Key Words: Aerial Image Object, Object Detection, Deep Learning, Weakly Supervised Learning

目录

图目录

表目录

第一章　绪论

1.1 课题背景和研究意义

随着航拍以及遥感技术的发展，人们能够利用成像设备获得大量的航拍视角的地面光学图像信息。高清航拍图像目标检测技术在这种背景下具备了一定的研究价值，精确的航拍目标检测算法能够对地面目标进行准确的识别与定位，并进行场景分析，该技术具有广泛的应用前景和研究价值。在军事领域，可以利用航拍图像检测地面的军事目标，如舰船，飞机等，并达到一定的军事目的；在民用领域，可以在城市上空放飞一些具有摄像功能的无人机，对城市进行实时火情监控，利用航拍目标检测技术，可以迅速的锁定火情的发生情况，并即时汇报给地面单位，以主动进行火警支援，这大大降低了火灾扩大的可能性，提高了救援速度。在高速公路上分路段放飞一些载有摄像设备的无人飞行器，利用航拍目标检测技术和高速计算设备对公路场景进行实时分析，这样，当公路上发生紧急情况的时候，有关部门即可在第一时间抵达现场，这将会大大降低因时间延误而造成的人员伤亡和财产损失。航拍场景下的目标检测技术还可以用于海洋漏油检测，灾后搜救，农作物病虫害检测等领域。

目前，空中摄影以及遥感技术已经比较成熟，人们借助飞机及卫星可以获取高清且稳定的地面视觉影像。机载摄像机高度低，操控灵活，可以采集到清晰的地面图像数据；星载光学设备则能获取大范围的光学图像。然而，由于航拍目标检测的技术难度较大，识别率高的航拍目标检测技术还有待开发。国外著名高校如南加州大学，康奈尔大学等都在积极开展基于航拍视角的图像理解技术的研究。美国军方也在研发一种全自动化的无人战机，运用目标识别技术，识别敌方目标，并达到军事目的。然而，国内的此类研究正处于起步阶段，仅有很少的高校在进行类似的研究，这些投入的力度是远远不够的。

近年来，目标检测技术作为航拍目标检测的母课题取得了很大的发展，人脸，车牌等目标检测技术走向了实际应用。航拍目标检测作为目标检测的子课题逐

渐引起注意，然而，航拍目标检测有一个特点：那便是它的视角变化很大，这也带来了一些难点，比如，航拍图像中，视角的变化导致目标的长宽比不一致，角度分散较广；航拍图像的分辨率较低，尤其是星载光学设备获取的图像，分辨率往往在 0.1 米以上；许多目标容易引起虚警，如建筑物的外形容易和汽车混淆等；航拍目标的标定耗费巨大的工作量，航拍范围大，目标多，标定困难。为了实现航拍目标的检测，学者们做了许多尝试，多种手工特征，如 Haar-like[1], HOG (Histogram of Oriented Gradient) [2-4],LBP(Local Binary Patterns)[5]被用来表示航拍目标，分类方法如 SVM(Support Vector Machine)[6]和 PLS(Partial Least Squares)[7]等被采用为航拍目标检测的分类器。这些特征和分类器推动了航拍目标检测技术的进步。然而，这些特征和分类器的组合难以适用于角度多变的航拍目标，人们不得不采用旋转原图的方式去检测航拍目标，这又带来了大量的计算量和虚警。

因此，如果能够提取一种角度不敏感的航拍特征或者提出一个效率更高的框架，将会大大推动航拍目标检测技术的发展。本文的目标就是基于高清晰度航拍图像中飞机和汽车的检测算法研究，设计角度不敏感特征算子，推动航拍目标检测算法的整体进步；基于目标检测框架的最新成果以及旋转敏感特征算子，设计航拍检测的非旋转原图框架。

另外，目前大多数目标检测的方法都是在建立在有监督的机器学习的基础上。监督学习是从标记的训练样本中训练出一个模型，每个训练样本都有输入对象和期望的输出信号（类别信息）组成。通过有限数量样本来调整分类器使其进行分类时产生的错误概率最小。这样的学习方法在目标检测领域已经有很大的进展。但是，监督学习往往伴随着大量的样本数据标定，需要耗费前期大量人力投入。而无监督学习用于处理未被分类标记的样本集。这两种传统的学习方法在高清航拍图像这种场景复杂，目标类型多样的情况下各有明显的不足。弱监督学习是模式识别和机器学习领域研究的热点问题，是监督学习与无监督学习相结合的一种学习方法。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的。

本文提出了采用弱监督学习的方式来自动标定目标，并基于弱监督学习标定的正例，训练分类器，该弱监督学习算法可以降低航拍目标标定的工作量。

本文受到了以下课题的资助:

（1）“危险化学品事故全过程遥测预警的关键科学问题研究”，国家 973 计划，2011~2015。

（2）“基于多源数据的飞行器进近威胁目标检测跟踪及行为预测”，国家自然科学基金重点项目，2011~2014。

1.2 国内外研究现状

一方面，航拍目标检测所采用的方法与技术多继承自经典目标检测算法如：车牌检测、人脸检测；另一方面航拍目标检测中提出新的特征、模型以及分类器也可以推广到其它目标的检测。因此目标检测技术和航拍目标检测技术的发展是相互促进和推动的。下面先从目标检测的角度介绍国内外的研究现状，再介绍航拍目标检测的内容。

常规的目标检测框架主要包括三方面的技术：一是感兴趣区域提取。感兴趣区域提取的目标是产生一些可能是目标的“可疑图像块”，又称感兴趣区域（ROIs），感兴趣区域提取算法的主要研究价值在于设计查全率高且准确率也较高的感兴趣区域提取算法。二是特征提取。即通过对感兴趣区域中图像信息的运算，将该区域的图像数据变换为一个能够准确描述该目标，并对光照，微小形变等因素不敏感的行向量。该特征应该具备较好的区分性和内聚性。区分性的含义是该目标的特征要和其它目标或者背景的特征有较大的距离，内聚性的含义是该目标的不同实体的特征要保持近似，或者说距离较近。三是分类器的构建。分类器是为了区分背景和目标，分类器的一般功能如下: 输入一个图像块的特征到分类器中，分类器判断该图像是否是待检测目标，是目标的概率有多大。这三方面的技术互相制约，相辅相成，共同决定着目标检测算法的效率。

在感兴趣区域选择算法方面，目前的研究方向逐渐从窗口扫描穷举法过渡到基于生物学先验知识的感兴趣区域提取算法。窗口扫描穷举法是最早的感兴

趣区域提取算法，该算法把图像中所有可能成为目标的图像块都当成感兴趣区域；由于目标的尺度不确定，通常要先把原图降采样和升采样到不同的尺度，然后再用一个窗口自左向右，自上向下地扫描，这个过程中窗口所覆盖的图像区域都作为感兴趣区域。这种方法会得到非常多的感兴趣区域，这给后续的特征提取和分类算法带来了巨大的计算量，通常一幅一百万像素图像将会产生数十万个候选框。由于窗口扫描穷举法输出的感兴趣区域过多，保证尽可能多地覆盖目标的情况下尽量减少感兴趣区域的数量就成为了一个研究方向。近年来，一些感兴趣区域选择算法被提出，2012 年，程明明等人发表的文章[8]中提出，在认知科学与神经科学中，人类能够在识别物体之前对物体有一个预判，人类只会对眼睛中的一小部分图像信息进行处理，而其他的大部分都被忽略。也就是说，人类的视觉机制中有一种快速筛选感兴趣区域的能力。该文章提出一种简单且高效的特征：BING（Binarized Normed Gradients for Objectness），来模仿人类的这种预选感兴趣区域的能力。程明明认为目标是一些具有确定边缘的独立物体，物体的边缘具有一定的共性。因此 BING 特征是一种梯度特征。这种特征简单而高效率，在许多实验中取得了良好的实验结果。文章[9]提出了一种基于图像分割的感兴趣区域提取算法。文章认为，物体由一些相邻的图像块组成。基于这个观点，该文首先中将图像分割为一些小的区域，然后将这些区域合并整合，这个过程中的所有区域都作为感兴趣区域。这个算法能够将感兴趣区域数量从数万降低到数千或者数百，并且能涵盖绝大多数目标。

在特征提取方面，首先要介绍 Haar-like 特征，该特征由 Papageorglou 和 Pogglode 等人在文章[1]中提出，该文章提出采用 Haar 小波描述行人，人脸和车辆。这种特征的优点是能够抓取显著区域，但是，该特征不太适合表示边缘轮廓信息，容易受到目标形态,光照条件及视角的影响，因此一般应用范围有限。在文献[2-4]中，作者提出使用 HOG 特征描述子表述人体，该描述子借鉴了 SIFT（Scale Invariant Feature Transform）[10][11]特征点中运用梯度方向直方图表示目标的思想。后来，在 HOG 特征的基础上，涌现出一些改进版本的特征[12]。Mu 等人提出使用改进的 LBP 算子来描述人体。在文献[2]中，作者为解决部分遮挡条

件下的人体检测问题，采用 HOG 特征和 LBP 特征相结合的方法。LBP 特征可以表述纹理，对单调灰度变化有敏感性；当背景比较复杂，特别是有干扰边缘时，HOG 特征将受到很大影响，而此时 LBP 特征可以滤除背景噪声。因此，HOG 特征与 LBP 特征相结合表示人体目标，可以取得较好的检测效果。以上特征是一些手工设计的特征，是掌握了本领域知识的人根据图像或者目标的特点去设计的特征。然而，目标的种类是非常多的，去为每一种目标设计相应的特征是非常困难的，也是不现实的。因此人们开始探索让计算机从图像中自动学习提取特征的方法，深度学习便是近年来兴起的智能学习型特征提取算法。深度学习是在研究传统 BP 神经网络的过程中提出的，它的基本原理是通过样本的基本信息学习目标的底层表示，再通过对底层表示的学习，形成目标的高层抽象表示。深度学习一般通过多层的神经网络实现。反向传播算法作为传统的训练多层网络的算法，对层数较多的网络训练不理想，因为反向传播过程中，容易造成局部最优解。 为了解决这个问题，Hinton 以及他的学生们经过多年的探索，提出逐层训练网络的初始化算法[41]，使得神经网络的训练不再陷入局部最优解。Lecun 等人提出卷积神经网络[14]，卷积神经网络改进了全连接的网络结构，采用卷积核扫描的方式有效降低了参数数量，提高了训练鲁棒性。人们使用卷积神经网络，设计了很多良好的目标检测模型。

在分类器的设计方面，大致可以分为两种研究方向：第一种是基于概率的方法（Probabilistic Method）[42]，另一种是基于判别的方法（Discriminative Method）[6]。目标检测研究的初期，研究者门多采用基于概率的方法，如模板匹配，Fisher 判别等。后来，由于判别算法训练简单，检测效果好，人们逐渐地转向判别算法。行人检测、车辆检测、人脸检测等基本采用了基于判别的方法。比较常用的基于判别的方法有支撑向量机（SVM），Adaboost 算法等，一些研究从理论上证明基于判别的方法要优于基于概率的方法[15]。SVM 算法通过最大化判别曲面和支持向量间的距离实现对目标的分类。考虑到样本的非线性分布，SVM 方法中提出使用核理论将样本投影到高维空间中，并将核理论中的内积形式运用到优化模型的对偶规划中，进而在高维空间中求解线性分类器。Adaboost[16-18]算法

是 Boosting 算法中典型的代表算法之一，它采用加权投票机制，设计多个弱分类器，并且将每个弱分类器看作一个投票委员，通过误差惩罚重采样原则，每次加大错分样本的权重以训练新分类器。理论上，Schapire 等人[19]已经证明 Adaboost 算法的迭代过程是收敛的，并且还证明了 Adaboost 算法也遵循边界最大化原则。

航拍目标和传统目标检测相比，有其独特的特点，那就是角度的多变性；一般的特征不具备角度敏感性。为了解决这个问题，传统的做法都是将待检测图像进行旋转，然后用穷举法选择候选框，最终再进行判别，这种做法计算量太大，虚警也会在候选框增多的过程中增多。如文章[20]，采用了几种常用特征进行组合，并训练了归一化角度的检测器。并通过旋转原图进行航拍目标中车辆的检测，该文所提出的算法的计算量很大。因此，研究角度敏感的的航拍目标特征表示方法以及无需旋转原图的航拍目标检测框架具有一定的价值。

1.3 本文研究内容

1. 本文提出使用高维数据可视化工具来描述高清航拍图像中目标的特征与角度的关系。利用这种方式，本文从卷积神经网络的特征分析着手，发现了卷积网络的某些层的特征层可作为角度不敏感特征描述子。

2. 基于角度不敏感特征描述子，本文提出无需旋转待检测图像的检测框架。采用基于分割的算法提取感兴趣区域，使用混合的卷积特征进行特征描述，使用支持向量机分类器对感兴趣区域进行准确分类。这个框架比原本的旋转原图的框架有了一定的改进。

3. 针对目标检测需要标定大量训练样本的问题，本文提出采用弱监督学习的方式进行正例挖掘，以传统的手工标定带来的人力负担。本文在航拍飞机数据集上做了正例挖掘实验，实验证明，本文的正例挖掘算法在高清航拍飞机数据集上表现良好。

1.4 本文的组织结构

第一章，绪论。介绍了航拍目标检测的研究背景和研究价值。介绍了当前的研究现状以及本文的主要研究内容，并对航拍目标检测的主要技术进行了初步的阐述和说明。

第二章，航拍目标检测的相关工作与技术。分析了目标检测常用的特征描述子以及常用的分类方法。将目标检测的框架拆分成三个部分：感兴趣区域提取研究，特征描述方法研究，分类器研究。并分别介绍了每个部分的代表性技术和优缺点，以及发展趋势。

第三章，监督式高清航拍目标检测算法研究。本章提出使用高维数据可视化工具来描述高清航拍图像中目标的特征与角度的关系。利用这种方式，从卷积神经网络的特征分析着手，发现了卷积网络的某些层的特征可作为角度敏感特征描述子。基于角度敏感特征描述子，本章提出了无需旋转待检测图像的检测框架。并做了定量的实验和分析。

第四章，弱监督高清航拍目标检测。针对目标检测需要标定大量训练样本的问题，本章节提出采用弱监督学习的方式进行正例挖掘，取代传统的手工标定。本文在航拍飞机数据集上做了正例挖掘实验，实验证明，本文的正例挖掘算法在高清航拍飞机数据集上表现良好。

第五章，总结与展望。总结了本文的主要工作，分析了论文的不足、列举了高清航拍目标检测仍然存在的难点和问题，展望了未来的研究方向。

第二章　目标检测相关工作与技术

近年来，航拍目标检测作为一个前景广阔的课题引起了广泛的关注。快速鲁棒的航拍目标检测算法有望在交通管制，紧急情况处理以及大规模图像内容分析方面获得应用。人们提出了一些的航拍目标检测算法[20]，对飞机以及汽车进行了检测。然而，角度多变性问题依然是一个挑战。大多数现存的特征表示以及检测方法无法解决这个问题。在大多数的航拍目标检测解决方案中，人们不得不采用固定角度的目标样本训练检测器，并对原图进行旋转检测[26]。

航拍目标检测是目标检测[26-30]的子课题，而目标检测的技术是相通的。目标检测的目的是让计算机代替人眼找出图像中的相关目标，并且在图像中标出目标的大小和位置，这个过程也可以称作目标提取。经过数年的发展，目标检测算法逐渐形成了一些固定的流程：先从图像中提取一些可能是目标的区域，也就是“感兴趣区域”；在感兴趣区域中提取特征；对特征进行判别并确定目标位置。因此目标检测的研究通常也分为感兴趣区域提取算法研究，特征描述方法研究，分类器研究三个方向。本章也将分别介绍这三个方面的相关技术。

2.1 感兴趣区域提取

在感兴趣区域选择算法方面，目前的研究方向逐渐从窗口扫描穷举法过渡到基于先验知识的感兴趣区域提取算法。先验知识有很多种，可以是生物学或神经科学的发现；也可以是对图像本身的一些研究。本章将首先介绍窗口扫描穷举法，然后介绍两种具有代表性的基于先验知识的感兴趣区域提取算法。

2.1.1 窗口扫描穷举法

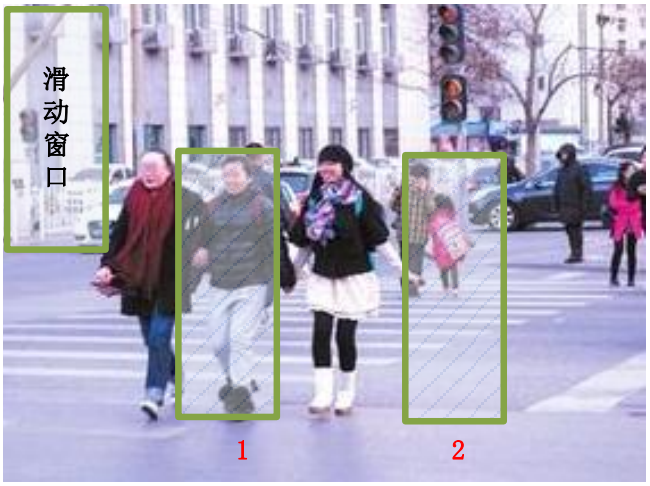

图 2-1 窗口扫描穷举法示意图

窗口扫描穷举法的基本思想是将所有可能成为目标的区域都作为感兴趣区域，不使用任何先验知识。因为目标的大小不确定，这种方法就得在不同的尺度上分别进行窗口的扫描并进行区域提取。如图 2-1 所示，以行人检测为例，窗口扫描穷举法用左上角的窗口从左向右自上而下地选取感兴趣区域。可见，在当前尺度下，可以准确找到 1 号行人以及与 1 号行人大小相当的行人，而 2 号行人是一个距离镜头较远的儿童，窗口中会包含过量的背景。因此要想准确包围 2 号行人，就得将该图像进行上采样，使得二号行人变大一些。

窗口扫描穷举法法的优点是全面涵盖了可能成为目标的不同尺度的图像块，不会漏掉目标。然而这种方法也有其固有的缺点：没有利用目标本身的特性，也没有模仿人类的生物学机理；计算量大，输出的感兴趣区域数量很多，当图像较大的时候且对实时性要求较高的时候，该方法难以完成任务。

2.1.2 BING

2012 年，程明明等人发表的文章[8]中提出，在认知科学与神经科学中，人类能够在识别物体之前对物体有一个预判，只会对映入眼睛中的部分图像进行处理，而其他的大部分都被忽略。也就是说，人类的视觉机制中有一种快速筛选感兴趣区域的能力。基于这个假设，该文章提出一种简单且高效的特征：BING，

来模仿人类的这种预选感兴趣区域的能力。程明明等人认为，目标是一些有相似边缘的独立物体，因此 BING 特征的设计基于简单梯度特征，这种特征简单而高效，在许多实验中取得了良好的实验结果。Bing 的特征提取过程如下：

1. 使用 1 维的模板[-1,0,+1]计算图像横向和纵向的梯度图，并计算梯度幅值。

2. 对图像的梯度幅值图进行尺度变换，归一化到 8x8 的尺度，获得归一化梯度特征（NG）。

3. 利用二进制估计算法[43, 44]，对 NG 特征进行重新估计，获得 BING 特征。

Bing 分类器的训练过程如下：

1. 提取正负样本的 BING 特征，输入到 Linear-SVM 中训练得到一个 SVM 分类器，将其归一化，作为级联分类器的第一级。

2. 使用 1 中得到的分类器搜索训练样本所在的原图，这个时候可以得到很多目标框，采用 NMS（Non-Maximum Suppression）抑制，然后根据一定的规则选择一些候选框框作为第二级分类器器的训练样本正例；使用这个分类器去搜索训练难样本，生成第二级训练样本负例。

3. 将第 2 步生成的正例和反例样本输入到 Linear-SVM 中训练得到一个分类器，作为级联分类器的第二级。

在选择感兴趣区域的时候，首先穷举感兴趣区域，然后提取 NG 特征，并转化为 BING 特征；利用训练获得的打分模型对所有感兴趣区域打分，并选出打分足够高的作为最终的感兴趣区域。该方法模拟了人眼识别目标的机理，具有生物学理论支撑。与窗口扫描穷举法相比，BING 有效地降低了感兴趣区域的数量，速度很快且准确率很高。该方法缺点是会漏掉一些奇异的目标，比如：蛇，绳子等。

2.1.3 Selective Search

J. R. R. Uijlings 等人在文章[9]中提出了一种基于图像分割的感兴趣区域提取算法，文中称之为 Selective Search。作者认为，不管什么目标，都由颜色块组成。那么先把图像分割成一些最基本的小色块，感兴趣的目标就必然会是这些

色块的排列组合。作者利用这个先验知识改进了穷举算法，有效地降低了感兴趣区域的数量。

Selective Search 基于图像自身的信息从底层开始分割并合并整合，可以锁定不同尺度的目标，减少了传统“扫窗”方式所带来的海量的感兴趣区域数量。另外，计算区域相似度的时候，可以采用颜色、大小、纹理等多种策略，适应不同目标和场景。Selective Search 可以用在树林，公路等“非聚团”目标的检测上。然而，由于 Selective Search 依赖于初始图像分割，因此，图像分割的准确程度对结果的影响很大。该算法的主要过程如算法 1 所示。

算法 1：Selective Search

输入：彩色图片

输出：物体位置的可能结果 L

1. 对输入图像进行分割，获得初始图像区域集合： $R=\{r_1,r_2,...r_n\}$。

2. 初始化相似度集合 $S=\phi$，初始化可能结果 $L=R$。

3. 计算区域之间的两辆相似度，（相似度的定义结合了位置和图像的特征[9]），将其添加到相似度集合 S 中。

4. 从相似度集合 S 中找出最大值所对应的两个区域 r_i， r_j，并将其放入 L 中；将 r_i 和 r_j 合并成为一个区域 r_t，从相似度集合以及中 R 除去与 r_i 和 r_j 相关的信息，计算 r_t 与其相邻区域的相似度，将其结果添加的到相似度集合 S 中。同时将新区域 r_t 添加到区域集合 R 中。4 过程循环进行多次。

2.2 特征描述

图像特征，可以定义为对图像原始信息的线性或非线性整合。根据不同的应用场景，图像的特征有其相应的含义或定义方法。在目标检测领域，理想的特征应该具有可区分性，内聚性、以及高效等特性；另外还需要能够对图像亮度变化、尺度变化、旋转和仿射变换等不敏感。

根据应用场景提取相应的特征描述目标是计算机视觉研究中的一个重要研

究内容。例如，图像匹配中，常用的特征就是图像中的特殊位置，比如建筑物的角点，这些局部特征通常被叫“关键点特征”或者“兴趣点”，或者还可以叫“角点”，它们通常用其位置点周围像素块的非线性描述来表示。另一类特征是“边缘”，基于边缘特征的特殊形态，可以对图像进行匹配。边缘特征还可以作为视频中出现物体边界和遮挡情况的指示器。另外，在描述物体的时候，人们往往关注物体亮度的变化，计算机视觉中常常使用图像的梯度去刻画这个特征。因此，图像特征大致可以分为三类：点状和区域状特征、边缘特征、线段（几何）特征。

在目标检测的发展历程中，涌现出一些意义重大的特征。其中，SIFT、HOG，Haar-like、LBP 等特征是比较有代表性的，这些特征是科学家们根据先验知识，进行手工设计获得的特征，本章会重点介绍；另外，也可以使用无监督的方式从样本中学习特征提取的模型，本章将重点介绍比较有代表性的深度学习算法。

2.2.1 Haar-like 特征

在早期的研究中，人们仅采用图像的光学强度信息（即图像每个像素的的 RGB 值）进行特征提取，这使得特征的计算开销很大。Papageorgiou 等人提出利用小波变换提取特征[45]，这是计算机视觉领域第一种实时的人脸检测特征。Viola 和 Jones 根据这个思想，设计了 Haar-like[22][46]特征。Haar-like 特征将图像块内的像素分成一些相邻的矩形方块（方块称为特征模板），计算这些矩形图像块内的像素灰度值之和；并计算这些灰度值之和的差值，差值作为该图像块的特征。

积分图是遍历一次图像求出图像中所有区域像素和的快速算法，该算法大大的提高了图像特征值计算的效率。积分图借鉴了动态规划的思想，从图像的左上角开始，将图像左侧以及上侧的像素的灰度之和保存于内存中，这样当使用某矩形区域的图像灰度之和时可采用该矩形对应的积分图中右下角值减去左上角的值。

Haar-like 特征反映了图像灰度变化的情况，如边缘，纹理。人的头发颜色会

比周围的皮肤颜色深一些，嘴唇两侧也会比面部颜色深一些，光线亮的地方的颜色会相对浅一些，这样的色彩变化可以用 Haar-like 特征来捕捉，因此 Haar-like 特征检测色彩变化明显的人脸效果较好，而检测色彩变化小的目标效果就比较一般。由于矩阵的结构比较简单，因此只能描述水平，垂直和对角的色彩变化。因此，要想检测更多的目标，需要根据目标的形态改变特征模板的形状和大小，这也给 Haar-like 特征的广泛应用带来了瓶颈。

2.2.2 SIFT 特征

SIFT 特征的全称为尺度不点特征描述子，在计算机视觉中应用广泛。目标跟踪，图像拼接与匹配等领域都要用到 SIFT 特征。在 SIFT 特征的基础上，人们也提出了很多目标检测特征。SIFT 特征由 David Lowe[11]教授提出，并开放源码。该算法通过多尺度极值运算求取图像中的特征点，并采用梯度统计信息鲁棒地描述该特征点。SIFT 特征独特的计算方法使得它具有角度敏感性，并且对图像亮度的变化以及视角的微小变化不敏感。

在图像中提取 SIFT 特征的步骤如下：

1. 构建尺度空间，对图像进行金字塔运算,获得原图像的多尺度映像。这个过程采用不同尺度的高斯卷积核对原图以及降采样和升采样后的图像进行卷积，以模仿视角的远近变化，最终对相邻图像进行差分运算。

2. 采用局部高斯法寻找感兴趣特征点。为了降低计算量，通常通过计算图像的 Dog（Difference of Gaussian）的极值点来近似。

3. 去除不良点。边界和低亮度区的点是不具有代表性的，因此采用 Harris corner 检测算子去除不好的特征点。

4. 为特征点选取主方向。并且按照这个方向进行进一步的计算，消除方向对特征的影响。

5. 生成 SIFT 算子，通过区域梯度方向统计得到 SIFT 特征的描述子。

SIFT 特征利用了尺度空间的理论，模拟了图像的多尺度属性，使得图像的特征对尺度不敏感。另外，SIFT 特征利用了像素间的相对灰度值大小，梯度信

息等因素，使得该特征对旋转，亮度变化保持敏感性，对视角，仿射变换等也保持一定的稳定性。因此，SIFT 特征在图像匹配领域应用广泛，在 SIFT 特征的基础上，人们提出了一些性能很好的目标检测特征[2,3,4]。

2.2.3 HOG 特征

方向梯度直方图特征(HOG)由 Dalal 和 Triggs 于 2005 年提出，并首先应用于人体检测。HOG 特征通过计算和统计图像局部区域的梯度方向直方图来构成特征。由于 HOG 特征是在局部区域内做的梯度统计，因此，HOG 特征对小形变不敏感。

HOG 特征结合 SVM 分类器被广泛应用于目标检测中，尤其在行人检测中获得了极大的成功。HOG 特征提取过程如下：

1. 将输入图像转化为灰度图像。

2. 对输入图像进行颜色空间的标准化；调节图像的对比度，降低图像局部的阴影和光照变化所造成的影响，同时抑制噪音的干扰。

3. 计算图像每个像素的方向梯度和幅值梯度。

4. 将图像划分成小 cell（例如 6*6 像素/cell）。

5. 对 cell 中的梯度进行直方图统计，形成每个 cell 的描述子。

6. 将相邻几个 cell 组成 block，并将 block 内所有的 cell 的特征描述子串联起来，得到该 block 的 HOG 特征描述子。

7. 将该图像内的所有 block 的 HOG 特征描述子串联成一个长向量，这个长向量就是该图的 HOG 特征。

HOG 特征独特的提取过程使得其具有很多优点。首先，HOG 特征提取的时候将原图划分成一些方格，并在方格上统计梯度，因此，HOG 提取的是一种局部统计信息，即使图像有一些微小的扭动或者光学变化，特征也可以保持基本不变。由于 HOG 特征提取过程中的分单元处理办法，使得图像局部像素点之间的拓扑关系得以保持，这就保留原图像更多的信息。然而，从 HOG 计算过程本文可以看出，HOG 对描述形态变化较大的目标存在明显不足。

2.2.3 深度学习

Deep Learning[31-33]的概念起源于人工神经网络的研究，多层感知机（MLP）就是一种深度网络结构。Deep Learning 通过学习组合底层特征形成高层抽象表示，以发现数据的特征表示。而反向传播算法作为传统的训练多层网络的算法，对层数较多的网络训练不理想，因为反向传播过程中，容易造成局部最优解。Hinton[33]等人提出了非监督贪心逐层训练算法，为解决深层结构相关的优化难题带来了转机。Lecun[49]等人提出卷积神经网络，卷积神经网络[14]摆脱了全连接的网络结构，采用卷积核扫描的方式有效降低了参数数量，提高了训练的鲁棒性，卷积神经网络在目标检测领域有许多成功的案例。

下文将首先介绍神经网络的概念和 BP（反向传播）训练算法，然后基于这个基础，过渡到卷积神经网络的介绍。

神经网络由神经元组成，神经元是科学家设计的模仿脑电波的一种数学变换，如图 2-2 所示：

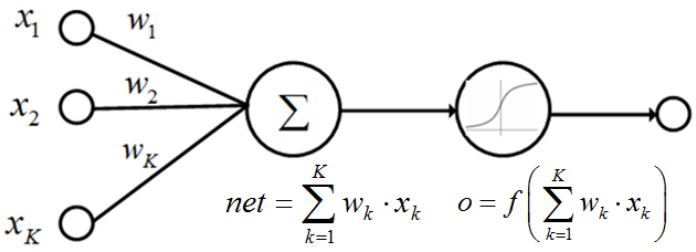

图 2-2 神经元示意图

该神经元用用数学公式来表达是：$o = f(\sum_{k=1}^{K} w_k * x_k)$，该神经元将一组输入神经元的数据 $\{x_1, x_2 ... x_k \,|\, x_i \in R, i = 1, 2...\}$ 转化为该神经元的输出数据 $o = f(\sum_{k=1}^{K} w_k * x_k)$，其中 f 为激励函数，一般选用 sigmoid 和正切函数等，w_k 为神经元之间的连接权。这就如同人脑的一个神经元，会接受其他一些单元的激励，然后再根据连接权的强度给出一定的反应。当然这个反应也会向传播到神经元网络中的其他神经元。

神经网络是由神经元组成的网络。实际应用中，我们使用的网络一般是单向非闭环网络。图 2-3 所示是一个典型的单向开环全连接神经网络，可见，输入数据 $\{x_1, x_2...x_k \,|\, x_i \in R, i = 1,2...k\}$ 经过隐藏层神经元向前传导，直至得到输出数据，也就是神经网络的输出 $\{o_1, o_2...o_m \,|\, o_i \in R, i = 1,2...m\}$。

我们一般采用后验误差最小化的原理来训练神经网络，也就是说，通过最小化网络输出和预期输出之间的误差的原则来搜索神经网络的参数空间（即神经网络的连接权），这个过程一般通过梯度下降法实现，而梯度的计算采用反向传播算法。神经网络训练的基本步骤如下：

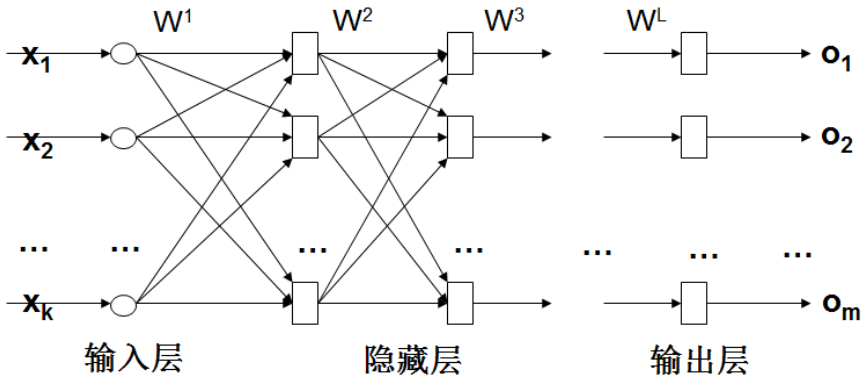

图 2-3　全连接 BP 神经网络

1. 随机初始化神经网络参数。

2. 前馈计算：根据样本集中的样本 $\{\vec{x}, \vec{y}\}$ 计算实际输出 $\vec{o}$ 和预期输出之间的误差 $E = 1/2 * \left\| \vec{y} - \vec{o} \right\|^2$。

3. 优化误差函数：通过反向传播算法来求解优化问题 $\min\limits_{W_1, W_2...} E = 1/2 * \left\| \vec{y} - \vec{o} \right\|^2$。

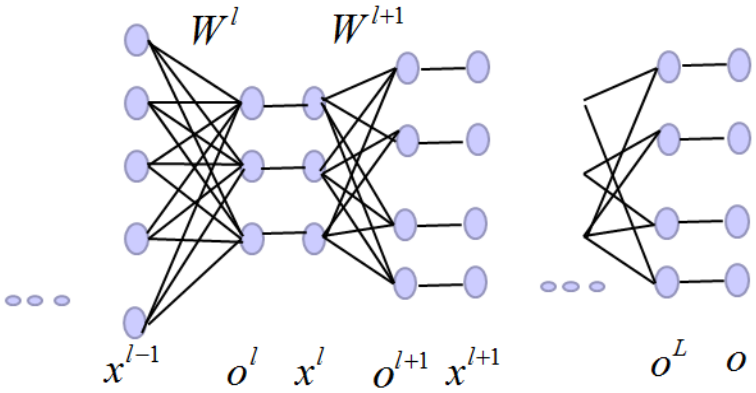

图 2-4 神经网络参数示意图

训练算法的初始化和前馈计算是显而易见的；关键在于优化的过程。优化过程采用梯度下降法实现，并因此需要计算梯度 $\frac{\partial E}{\partial w^l}$。梯度的计算采用反响传播算法。下面推导反向传播算法：

如图 2-4 所示神经网络有公式 2-1 和 2-2 的正向传递关系：

$$o^l = W^l x^{l-1} + w_0 \tag{2-1}$$

$$x^i = f(o^l) \tag{2-2}$$

那么根据链式法则，有 2-3 的关系：

$$\frac{\partial E}{\partial W^l} = \frac{\partial E}{\partial o^l} * \frac{\partial o^l}{\partial W^l} = \frac{\partial E}{\partial o^l} * \frac{\partial o^l}{\partial W^l} = x^{l-1}(\delta^l)^T \tag{2-3}$$

其中：

$$\delta^l = \frac{\partial E}{\partial o^l} = \frac{\partial E}{\partial o^{l+1}} * \frac{\partial o^{l+1}}{\partial o^l} = \delta^{l+1} * W^{l+1} \cdot f'(o^l) \tag{2-4}$$

且：

$$\delta^L = f'(o^L) * \sum_{i=1}^{N}(y^{(i)} - o^{(i)}) \tag{2-5}$$

由此迭代地由后向前求出偏导数，就可以得到 $\frac{\partial E}{\partial w^l}$。

以上的网络是“全连接”的设计。从计算的角度来讲，相对较小的图像，从整

幅图像中全连接地计算特征并训练网络是可行的。但是，如果是更大的图像，通过这种全连接网络来学习整幅图像上的特征，将变得非常耗时。卷积神经网络是一种局部连接的网络。卷积网络的由分层二维排布的神经元组成，两层的神经元之间通过连接权进行连接。可以将连接权看作一个二维的模板，这个二维模板以卷积的方式与输入层数据进行乘积加和运算，并将结果输入 sigmoid 激励函数，得到输出层。多个模板就会得到多个输出层。如图 2-5 所示，左侧是输入层，右侧是输出层，左下角是四个 3×3 的模板，四个模板分别与原图进行扫描运算，得到四个相应的输出层。可见，这个复杂的网络仅有 3×3×4=36 个参数，而如果采用全连接的方式则需要数十万个参数。通过前向多层卷积连接，可实现性能良好的卷积神经网络结构。

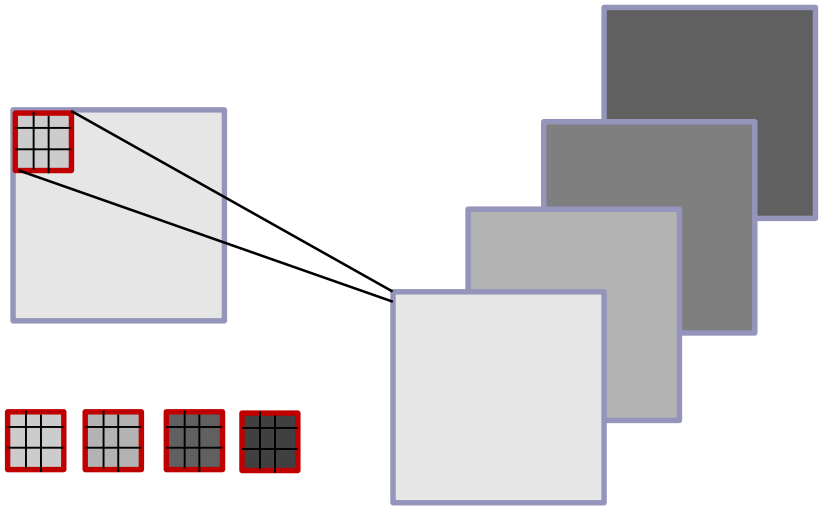

图 2-5 卷积网络示意图

实验证明，经过对大量图像的学习之后，卷积核能够捕捉一些图像的特征。前面几层一般捕捉边缘的细节特征，而后面几层则可以捕捉鼻子，眼睛等构件。在诸多图像识别的任务重，卷积网络表现出非常好的性能[48]。

2.3 判别算法

判别算法又称分类算法，分类算法是数据挖掘和模式识别中的一种非常重要的算法。分类的概念是指将目标映射到给定类别，这个过程一般通过分类器实

现分类。通常的做法是先标定一些数据，然后以人为规定的模型去学习数据的规律，该规律被刻画为模型的参数，最终使得该模型能够对样本进行分类。分类器是对样本进行自动分类的方法的统称，包括决策树、逻辑回归、朴素贝叶斯、神经网络，支持向量机，Adaboost 等算法。在目标检测领域，目前使用最广泛的分类算法为支持向量机和 Adaboost 算法，下文本文将重点介绍这两种算法的基本原理。

2.3.1 支持向量机

支持向量机[6]是一种二类的分类器，其基本模型是定义在特征空间上的间隔最大线性分类器。通常，SVM 分类问题被等同于寻找最优分类面。最优分类面不但要将两类样本正确的分开，而且要使分类间隔最大。如图 2-6 所示，由图中容易看出，当直线 H 作为分类线的时候，H1 和 H2 之间的间隔能达到最大，同时正反例距离分界面的距离也达到最大。SVM 分类器就是要寻找这样的分类面，而图中实心样本就是支持向量。支持向量机的学习策略就是间隔最大化，间隔最大化问题可以通过数学变换形式化为凸二次规划问题，并采用对偶原理进行参数求解。

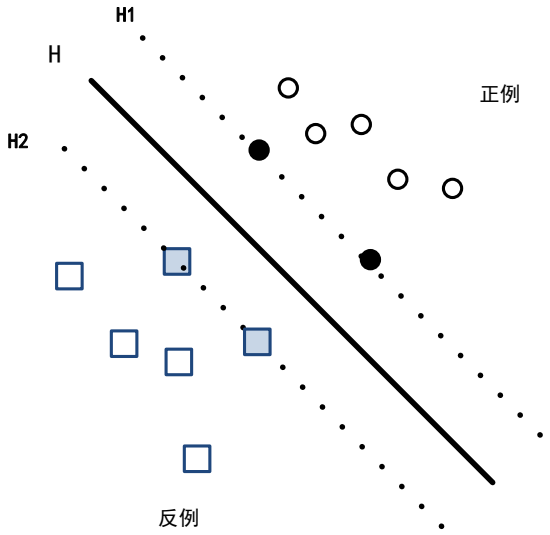

图 2-6 支持向量机原理图

支持向量机的类型包括线性可分支持向量机（linear support vector machine in

linearly separable case），线性支持向量机(Linear support vector machine)以及非线性支持向量机（Non-linear support vector machine）。当数据线性可分时，只需要采用线性可分支持向量机；而数据近似可分的时候一般采用软间隔最大化算法学习非线性支持向量机；数据不可分的时候就要采用核技巧学习非线性支持向量机。

支持向量机解决分类问题是在特征空间中进行的，一般先对欧氏空间的原始数据进行非线性变换，求得数据的特征，然后对特征空间内的全新数据集进行标定，最终将特征空间内的数据集分成训练集和测试集，并且进行模型训练，求得合适的参数。如果数据少，则可以进行交叉验证。

支持向量机是基于分界面的一种学习模型，这种分界面就是判别界面，如果训练数据维度低于三维，分界面退化为判别线，如果高于三维，就是判别超平面。假设给定一个特征空间上的训练数据集$T=\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$；其中$x_i\in R^n, y_i\in\{+1,-1\}, i=1,2...N$。并且，首先假设数据集线性可分。线性可分支持向量机的学习目标就是求得一组参数$\{w,b\}$，其中$\vec{w}$是和训练样本维度相同的向量，使得$\vec{w}*x+b=0$这个超平面能够将正负样本分在平面的两边。并且使得样本到超平面的距离最大。那么，一般情况将这个问题表示为如下的硬间隔最大化优化问题。其中 D 是支持向量与分界面的距离：

$$\begin{aligned}&\max_{\vec{w},b}\ D\\&s.t.\ y_i(\frac{w}{\|w\|}*x_i+\frac{b}{\|w\|})\geq D, i=1,2,...N\end{aligned}\tag{2-6}$$

即最大化超平面与训练数据之间的最小距离。通过一些数学变换，本文可以将上述优化问题改为：

$$\begin{aligned}&\min_{\vec{w},b}\ \|w\|^2\\&s.t.\ y_i(\vec{w}*x_i+b)-1\geq 0, i=1,2,...N\end{aligned}\tag{2-7}$$

这是一个凸二次优化问题，可以采用对偶原理进行转化，并用梯度下降等优化方法求解。

线性支持向量机是在线性可分支持向量机的基础上改进的。目的是为了解决这样的数据集：大部分数据线性可分，而部分数据线性不可分。解决方案便是修改约束函数，使得硬间隔变为软间隔：即在约束条件中加入代价因子 $\eta > 0$，使得约束条件变为： $y_i(\vec{w}* x_i + b) \geq 1 - \eta$

凸二次规划问题将会变为如下问题：

$$\begin{aligned} &\min_{\vec{w},b} \|w\|^2 + C\sum_{i=1}^{N}\eta_i \\ &s.t.\ y_i(\vec{w}* x_i + b) \geq 1 - \eta_i, i = 1,2,...N \\ &\eta_i \geq 0,\ i = 1,2,...N \end{aligned} \tag{2-8}$$

对于数据大部分线性不可分的问题，一般采用核变换对数据进行变换，在变换后的空间内训练支持向量机，这也称非线性支持向量机。

如图 2-7，矩形和圆形分别表示正，负样本数据。左边是原空间，可见，唯有采用圆形的判别曲面才能正确的进行分类；经过核变换后，数据转变成了右边的模式，只需要采用线性分类器就可以进行正确的判别了，也就是说，在新

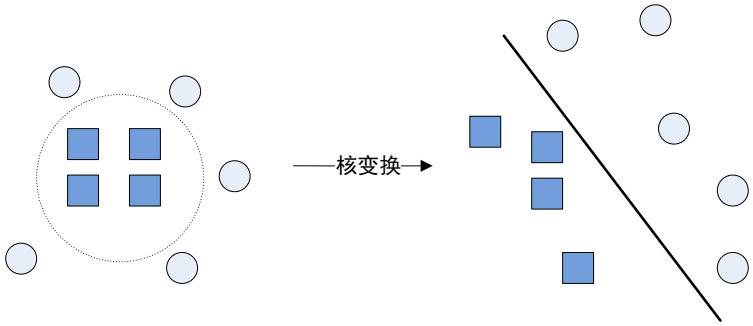

图 2-7 核变换示意图

的空间中，采用线性支持向量机就可以解决在原空间中线性分类器无法解决的分类问题了。这里其实采用的是一种划归的思想，将不便解决的问题转化为可以解决的问题。常用的核函数有多项式核函数，核函数，字符串和函数等，分

别适用于不同的数据分布，工程应用者可以自行选择合适的核。

2.3.2 Adaboost

Adaboost 算法是一种迭代算法，也是 Boosting 系列算法的一种。它的基本思想是针对同一个训练集进行训练样本权重的变化以获得多个训练集。使用这些训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。算法的核心原理是通过改变数据分布，它根据训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来更新每个样本的权值以及分类器的权重，分类错误的样本会被重点考虑，即增大权重。最后将每次训练得到的分类器用加权投票的方式融合起来，作为最后的决策分类器。Adaboost 算法提供的是一种框架，是组合弱分类器为强分类器的一种方法，而弱分类器的选择则是可以根据应用场景的不同而变化。Adaboost 算法在人脸检测等领域取得很大的成功，并广泛应用于计算机视觉领域。

2.4 本章小结

本章从目标检测的角度介绍了航拍目标检测常用的一些基本算法和重要理论。本文认为，目标检测的理论可基本分为感兴趣区域提取，特征描述，分类器设计三个方面。本章分别介绍了这三个方面的一些代表性算法并分析了各个算法的优缺点。

感兴趣区域提取方面，本文介绍了多尺度扫窗穷举算法，Bing 以及 Selective Search 三种代表性算法。特征描述方面，本文介绍了 HOG，SIFT，Haar-like 以及深度特征这三种代表性的特征。分类器方面，本文介绍了基于判别的 SVM 以及 Adaboost 算法。

第三章　监督式高清航拍目标检测算法研究

高清航拍场景下的目标检测技术应用前景广阔。然而由于航拍目标姿态多变等原因，目前没有合适的算法能够很好的处理这个问题。本章提出使用t-SNE[21]可视化工具分析卷积特征和角度之间的关系，并设计了角度不敏感混合卷积特征。

采用角度不敏感特征提取样本的旋转敏感特征，就可以进行无需旋转原图的高清航拍检测了。我们将角度不敏感特征融入 RCNN 框架[23]。并检测航拍中的飞机和汽车两种目标，实验证明，本文的算法性能良好。

该章节先分析航拍目标检测的难点，然后提出角度不敏感特征和检测流程，最后展示在高清航拍数据集中的飞机和汽车上的检测实验和结论。

3.1 难点分析与目标特性分析：

检测航拍图像中的目标，主要有以下一些难点：

1．航拍目标的角度变化范围大。

如图 3-1 所示，图像中飞机角度范围变化很大。角度不同的目标的特征会不一致，一般的特征和分类器难以处理旋转角度如此大的目标；如果采用一般的特征，将难以训练出合理的目标检测器。

图 3-1 航拍目标的角度多变性

2．航拍目标长宽比不定。

因为航拍目标角度不定，长宽比随着物体在图像中的角度的不同，会有很大的不同。长宽比范围太大，则不能采用多尺度扫窗穷举算法进行感兴区域选择，本文需要寻找使用于长宽比不定的目标的感兴趣区域提取算法。

3．航拍图像分辨率高，对检测速度要求高。

航拍图片一般比较大，要想快速的分析航拍目标，需要很快的检测速度。如果多尺度扫窗穷举法选择感兴趣区域，则难以进行实时分析。如果图像的尺寸为： $width = a, height = b$ ， 窗口尺寸为 $width = m, height = n$,扫描步长为 p, q，那么，那么这一幅图像的感兴趣区域的总个数为：$(\left\lfloor \frac{a-m}{p} \right\rfloor + 1) * (\left\lfloor \frac{b-n}{q} \right\rfloor$;

如果以步长（2，2）窗口大小为（36，36）去扫（1000，1000）的图像，那么就要得到的感兴趣区域总个数为：

$$(\left\lfloor \frac{1000-36}{2} \right\rfloor + 1) * (\left\lfloor \frac{1000-36}{2} \right\rfloor + 1) = 231360$$

这么多的感兴趣区域个数会给后续的特征提取与分类带来大量的计算压力。

4．当图像分辨率不够高时，目标会显得很模糊，这给本文的检测比较小的目标时候的目标检测算法带来了较大的挑战，如图 3-2 中的汽车目标。

图 3-2 航拍图像分辨率低导致汽车目标模糊

5．形态以及场景的多变性。

卫星遥感图像地貌十分复杂，包含了山地、草地、丘陵、沼泽、平原等。

图 3-3 建筑物容易造成虚警

多种复杂地面情况。飞机目标一般都停泊在机场，而机场又根据不同的使用需要建设在不同的地区。比如农用机场一般都建立在离农田不远的地方，而机场环境肯定会受到周围农田环境的影响，包括农作物、杂草等。民用机场一般建立在市区或离市区不远的郊区，这样可以节省时间和方便旅客的旅程。在市区的机场，机场环境的影响因素包括市区的交通工具、飞机机棚、街道建筑、汽车等。由于城市的交通工具比较多，比如轿车、卡车、客车等常用交通工具，为了方便旅客的出行，经常停在机场内，因而会对飞机的识别造成一定的干扰。除此之外，机场周围较大的建筑也是一种重要的影响因素。卫星从地球上空拍摄机场，机场建筑物会变成和飞机大小类似的目标，此时十分容易产生干扰现象。军用机场一般建立在山区、郊区或居民较少的沙漠地区，既有保密的需要，又有军事实际任务用途的考虑。建立在山区的军用机场，环境十分复杂。山区地势较高，拍摄的遥感图像存在一定的倾斜，而且周围有着大量的森林草木，都会掩盖机场的环境特征。建立在沙漠的军用机场，虽然环境背景较为单一，但是机场边缘和沙漠不易区分，从而使机场上的飞机目标识别有一定困难。正是由于机场环境的复杂性和多样性，使得飞机目标特征的实际提取过程变得较为困难。

飞机目标种类繁多，不同种类的飞机大小、颜色、形状等特征各不相同，因而提取的飞机特征区别也很大。但是，不管何种类型的飞机，一般都包括机头、

机翼和机身这几个主要组成部分。机翼的主要功能是为飞机提供升力，以支持飞机在空中飞行。为了维持飞机的稳定，一般副翼、襟翼和尾翼也安装在机翼上。另外，机翼上还安装了发动机、起落架和油箱等部件。机翼的形状和数量根据飞机类型而设定，但大都是流线型，这样有助于飞机在高空中飞行。机头的形状也是根据飞机的用途而设定的，比如客机和运输机的机头呈椭圆形，战斗机的机头呈尖针型。当然还有一些军用飞机的机头呈特殊形状，如美国的隐形轰炸机就呈三角形。至于飞机的机身，除了特殊用途的飞机，机身大都是相同的形状。

飞机的形状大体上是相似的，但是不同类型的飞机大小、颜色、形状又各有区别，这对特征的提取也有很大的影响。因此在提取飞机特征的过程中，必须要考虑这些情况。此外，遥感图像拍摄的飞机目标存在一定的几何变化，如平移、旋转、缩放等变化。不同的特征对飞机目标几何变化的敏感程度不同，因此几何变化也会对提取的飞机特征性能产生一定的影响。

而对于汽车目标，除了上述难点之外，就是该目标在航拍中的尺寸较小，另外由于汽车的方形结构和建筑物以及集装箱比较相似，如图 3-2，这两种物体在检测中会带来虚警。

3.2 监督式高清航拍目标检测算法

由于传统的检测算子难以适应目标的大幅度旋转，因此，科研工作者往往只训练单一角度的目标检测器，他们选择单一角度的汽车，然后训练检测器，这种检测器只能够检测这种角度的目标。后续只能采用旋转原图的方式进行其他角度目标的检测。每旋转一次只检测一种角度的目标。最终将各个角度下检测到的目标进行合并与筛选，这带来了大量的计算量。而如果能有一种特征，对角度不敏感，那么就不用这么做了。本文成功地用卷积神经网络提取出了角度不敏感特征。

基于角度不敏感特征，我们采用了图 3-4 所示的检测框架。先进行粗定位，

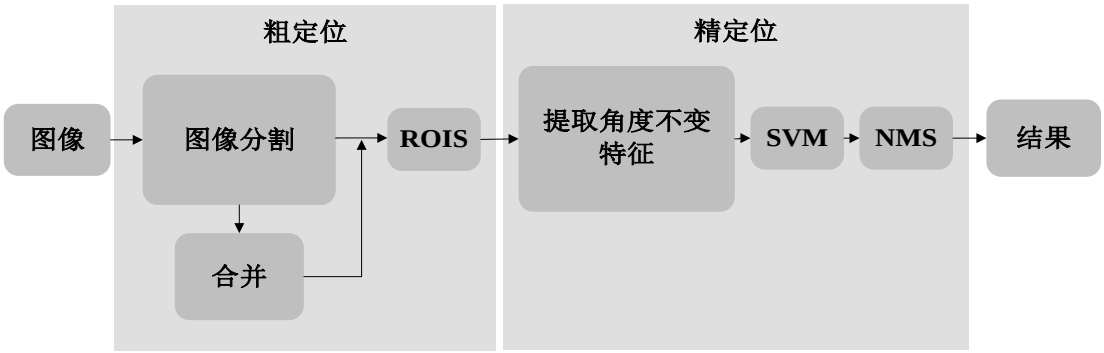

图 3-4 本文提出的检测框架

将疑似目标找出，然后再进行精确检测，去除非目标的感兴趣区域，最终得到检测结果。粗定位主要用于产生感兴趣区域（ROIs），感兴趣区域里面包含了待检测目标和一些背景。本文采用了 Selective Search 算法提取感兴趣区域，此算法将物体看成是一些色块或者是相邻色块的组合，该算法的第一步是分割出单一色块作为一组感兴趣区域，第二步根据位置信息和颜色信息等对相邻色块进行组合，合成新的感兴趣区域。精确检测部分主要用来剔除粗定位结果中的非目标图像区域，并给出最终的检测结果。本文对粗定位结果中的感兴趣区域计算角度不敏感特征，并送入训练好的 SVM 分类器算出相应得分，通过非极大值抑制求出最终的结果。为了方便，本文在 RCNN 的开源框架中进行实验。

3.2.1 感兴趣区域提取-粗定位

在航拍图像上进行感兴趣区域，需要考虑如下几个问题：

1. 适应不同尺度。由于航拍的时候成像设备会随着载体的高度变化，因此目标的尺度也会随之变化，因此航拍目标检测算法中的感兴趣区域提取部分要求必须能够提取不同尺度的区域。

2. 多样化。由于航拍中目标较多，基于可扩展性的考虑，要求感兴趣区域提取算法能够适用于多种目标。

3. 速度要快。航拍图像较大，因此对速度要求较高。

4. 输出感兴趣区域少。因为我们采用了卷积特征，计算量较大，因此不能有大量的感兴趣区域。

传统扫窗框架会产生上百万个感兴趣区域。那么特征提取算法和分类算法就将执行百万次，这需要耗费大量的计算资源以及时间。扫窗框架还要求目标的长宽比确定，而航拍中由于目标角度变化较大，长宽比的变化也相应较大，因此航拍目标检测不能使用扫窗框架。

为了减少窗口数量，一种基于分割算法的感兴趣区域提取算法 Selective Search 被提出，通过对图像区域的初始分割以及合并整合，将待选框的数量降低到数千或者数百，并且也不会遗漏目标。这种方法非常适用于航拍问题。Selective Search 采用图像分割有效地降低了感兴趣区域的数量，并且使用层次合并算法有效地使用于多尺度问题。Selective Search 采用了颜色、纹理、大小等多种策略对区域进行合并，因此能够适用于多种目标。

如图 3-5 所示，从左边到右边依次是从小型图像块到大型图像块的合并过程。图像块的矩形包络为感兴趣区域，可见，这种基于图像分割的算法，对长宽比不一致的目标能够正确的锁定。本文做过一定得测试，发现航拍数据集中绝大多数的目标会包括在 Selective Search 算法输出的结果中。Selective Search 结合了穷举法和图像分割。它相比穷举法能充分利用先验信息方法有效减少位置数量。它是一种基于图像解析学原理的选择性搜索。Selective Search 的基于图像自身的信息从底层开始分割并合并整合，可以锁定不同尺度的目标，省去了传统“扫窗”方式所带来的海量的窗口数量。另外，计算区域相似度的时候，可以采用颜色、大小、纹理等多种策略，适应不同目标和场景。Selective Search 可以用在树林，公路等“非聚团”目标的检测上。然而，由于 Selective Search 依赖于初始图像分割，因此，初始图像分割的准确程度对结果的影响较大。

基于 Selective Search 的优点，本文采用 Selective Search 作为感兴趣区域提取的方案。另外，我们利用先验知识过滤掉 Selective Search 的结果中长宽比不合理，以及长与宽的大小在合理范围之外的区域。

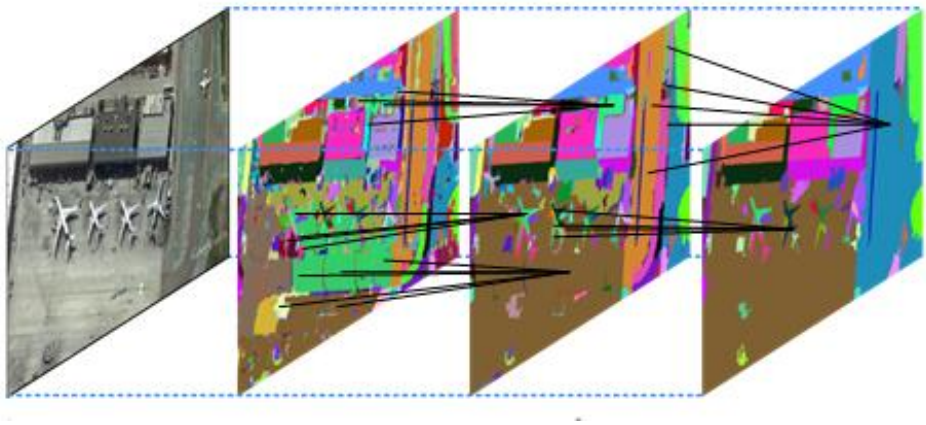

图 3-5 Selective Search 在航拍数据集上的示意图

3.2.2 特征提取-卷积特征与混合卷积特征的角度敏感性分析

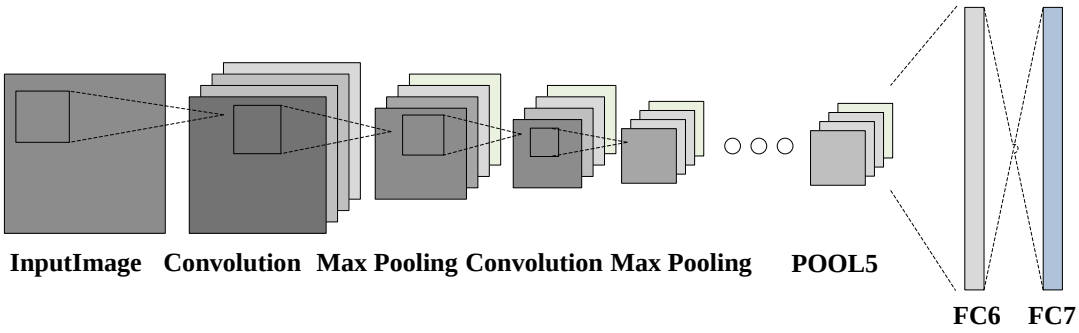

图 3-6 AlexNet 深度卷积神经网络

受 Hinton[33]用卷积神经网进行具有较大角度变化的文字进行识别的启发，本文采用深度卷积网来尝试进行角度鲁棒特征提取。为了实验能够快速进行，本文采用经过 Imagenet[34]训练的 AlexNet 网络进行角度鲁棒特征提取的尝试，并采用 t-SNE[21]作为特征选择辅助工具。

如图 3-6 所示，AlexNet 卷积网络包含五层卷积层，其中，三个卷积层后面都紧跟着一个池化层，另外，在这八层后面还堆叠着两个池化层，最后，两个全连接紧随其后。最新的工作说明，通过大量数据集训练的深度卷积网，能够被泛化至其他的视觉任务。另外，网络不同层针对不同的数据集有不同的表现。比如说，在 PASCAl VOC[21]目标检测挑战中，FC7 表现最好，然而，在场景识别任务中，FC6[39]表现最好。在本文的工作中，本文综合考虑 POO5，FC6，FC7

以及它们的组合来进行角度鲁棒特征提取，并寻找对角度不敏感的组合。

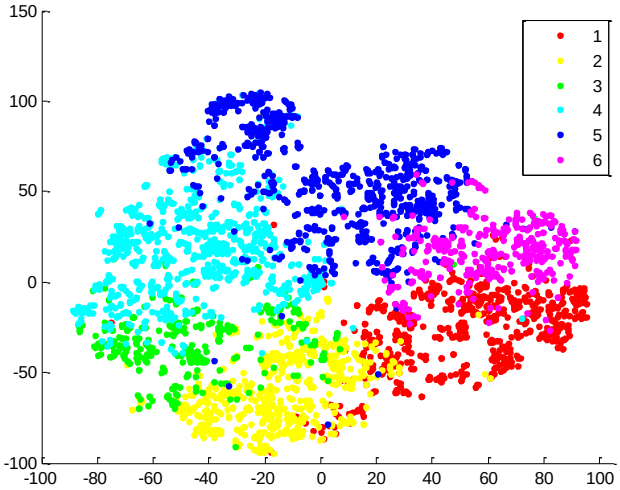

图 3-7 飞机的 POOL5 特征和角度的关系，相同角度（颜色）有明显聚团

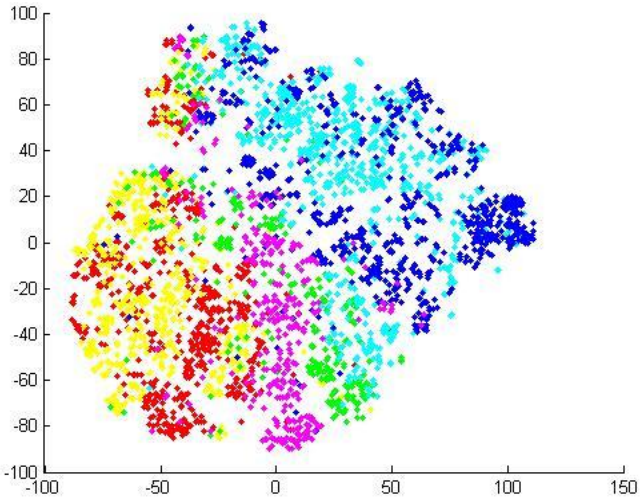

图 3-8 飞机的 FC6 特征和角度的关系，相同角度（颜色）无明显聚团

本文采用 t-SNE[21]作为可视化工具来进行提取，t-SNE 是一种可视化高维数据的技术，它将高维度数据映射到二维或者三维空间中，这样就可以直观地看出高维数据之间的距离关系。

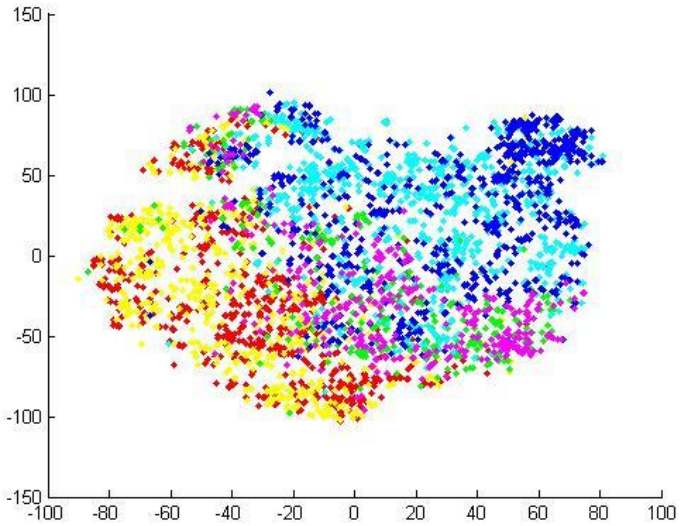

图 3-9 飞机的 FC7 特征和角度的关系，相同角度（颜色）无明显聚团

为了研究卷积特征对角度的鲁棒程度，本文尝试去分析特征的和角度的关系。我们用 t-SNE 将卷积特征的维度降低至二维，并将相邻角度的目标标记成相同颜色绘制在图中，如图 3-7，3-8 和 3-9，每个点代表一个样本的特征在二维空间的映射，每 60 度一个颜色，依次为 0-60 度，61-120 度，121-180 度，181-240 度，241-300 度，301-360 度。

从图 3-7 中可以看出，在 POOL5 空间中，相邻角度的飞机自然地聚成了一团，然而，在 FC6 和 FC7 空间中却没有出现明显的聚团。因此我们推断，POOL5 能够对旋转因素很好的建模。而 FC6 和 FC7 可能是对其他的因素建模了。理论上说，特征选择的过程是扎根于 Disentangling[40] 特征学习的，即用不同的特征去对不同的因素进行建模是非常合适的。既然 POOL5 能够对特征进行很好的建模，相信 POOL5 在旋转目标的检测中会有相对较好的表现。另外，通过分析，我们认为 FC6，FC7 应该是对其他的因素进行建模，比如说色彩，长宽比等，本文尝试将 POOL5 和 FC6，FC7 结合起来，并获得了性能上的提高。

另外，本文用基于 ACF（Aggregate Channel Features）[23]特征的检测器作为对比实验。ACF 采用 HOG，颜色特征，以及梯度特征进行多通道组合，是很多检测任务中性能最好的检测器之一。

3.2.3 分类器-线性 SVM

对于感兴趣区域提取阶段产生的感兴趣区域，本文以 3.2.2 中讲述的方法提取特征，并且训练了一个线性 SVM 分类器。采用线性 SVM 分类其的原因是混合 DCNN 特征维度高达一万多维，线性 SVM 计算量较小。分类结束后会有一个非极大值抑制的过程来抑制一些不精确正例，并获得最终精确检测结果。

3.3 实验

3.3.1 数据集

参考 Caltech[24]数据集的标定方式，我们组织人力在 GoogleEarth 上标定了 SDL 高清航拍数据集，分别为飞机数据集和汽车数据集，数据集的明细如表 3-1：

表 3-1 SDL 高清航拍数据集明细表

	飞机图像	飞机样本	汽车图像	汽车样本	反例图像
Version 1.0	600	3591	310	4475	492
Version 2.0	400	3891	200	2639	408
合计	1000	7482	510	7114	910

数据集共两个版本，分别为 Version 1.0 和 Version 2.0，总计耗费人力为 10 人*5 周。

数据集中，共含有飞机的图像 1000 张，包括飞机 7482 架；标定的结果按行保存在 txt 格式的文件中：例如：P0021.txt 中共 9 行，每行为 13 列；那么就说明 P0021.png 中有 31 架飞机，每列的属性分别为：

$$(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4, theta, x, y, width, height)$$

其中， $x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4$ 为飞机的矩形包络， $x1, y1$ 为飞机左上角坐标，从左上角开始，顺时针依次为 $x_2, y_2, x_3, y_3, x_4, y_4$， $theta$ 为机尾指向机头的向量与 x 轴正向的夹角； $x, y, width, height$ 为飞机的 Bounding Box。

汽车数据集共有含有汽车的图像 510 张；共含有汽车 7114 架；标定的结果按行保存在 txt 格式的文件：例如：P0310.txt 中共 26 行，每行为 13 列；那么就

说明 P0310.png 中共有 26 架汽车，每列的属性分别为：

$$(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4, theta, x, y, width, height)$$

其中 $x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4$，为汽车的矩形包络，坐标点顺时针排列，但起点不定； *theta* 为车头车尾的连线和 x 轴的夹角，0-180 度之间。 $x, y, width, height$ 为汽车的 Bounding Box。

3.3.2 实验方法与结果

通过实验，我们测试了单层卷积特征和混合卷积特征的表现并结合 t-SNE 进行了分析。实验是在 SDL 航拍高清数据集的 Version 1.0 上进行的。这个数据集包括飞机数据集和车辆数据集，其中，飞机数据集包括 600 幅图片，其中有 3591 个飞机；汽车数据集包括 310 幅图片，其中有 4475 个汽车。这些样本按是角度均衡地筛选的，通过图 3-10 可以看出，本文的样本角度非常均衡。每个集合都被分为训练集和测试集，分别为（500，100），（250，60）。

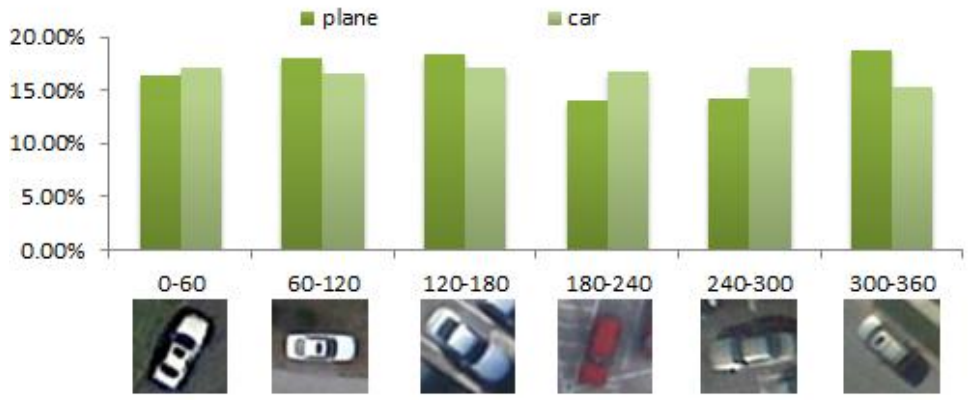

图 3-10 样本的角度分布直方图

目标检测算法将会得到一个带着 SVM 得分的包围框列表，性能评估便是在这个列表和标定列表上进行的。要判断目标检测算法的精度，首先要判断检测结果是否能够和原图象上的真实目标所重合，那么就得有一个标准来衡量重合的程度，然后再设定一个阈值，当重合率大于这个阈值时，就认为目标被检测上。

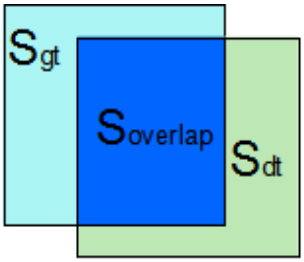

图 3-11 检测结果判定准则示意图

如图 3-11 所示: S_{dt} 为检测结果位置；S_{gt} 为标定的位置；$S_{overlap}$ 为二者的交集，即：

$$S_{overlap} = S_{dt} \cap S_{gt} \tag{3-1}$$

当检测结果和标定结果满足以下准则时，认为目标被正确检测到：

$$a = \frac{S_{dt} \cap S_{gt}}{S_{dt} \cup S_{gt}} = \frac{S_{overlap}}{S_{dt} + S_{gt} - S_{overlap}} \geq 0.5 \tag{3-2}$$

另外，检测性能的判断是对大量图像检测的统计结果。本文采用全部待检测目标的检测准确率（ *precision* ）和召回率（ *recall* ）来综合评判检测的准确程度。

准确率指将非目标检测为目标的百分比，定义如下：

$$precision = \frac{true\ positives}{false\ positives + true\ positives} \tag{3-3}$$

召回率的定义如下：

$$recall = \frac{true\ positives}{true\ positives + \ false negitives} \tag{3-4}$$

3.3.3 实验结论分析

由表 3-2 以及图 3-12 至 3-13 可以看出，DCNN 特征的表现强于 ACF。对于 DCNN 的单层特征来说，汽车的 FC7 比 POOL5 和 POOL6 表现的更好，然而对于飞机来说，POOL5 表现的最好。另外，混合的特征可以显著提高性能，其中包含 POOL5 的混合特征性能最好。这也从实验的角度上证明了 POOL5 能够对

旋转比较好的建模的观点；另外，FC6 和 FC7 层的特征提供了互补的信息。

图 3-14 展示了一些检测结果示意图，可以看出，绝大部分的目标能够被正确的检测，仅有几个目标被漏检：当目标与背景颜色比较相似的时候容易被漏检，通过实验验证，是图像分割的时候漏掉了该目标，图像分割算法容易将背景和目标看作一个基础色块。另外，当目标被遮挡较大的时候会出现漏检。

综上所述，混合深度卷积特征能够进行对角度不敏感的目标检测。拥有这些特征，就不必要去采用旋转检测的流程，这大大的降低了航拍目标检测的工作量。

表 3-2 检测结果

特征	汽车（recall=0.8）	飞机（recall=0.9）
ACF(baseline)	0.542	0.511
POOL5	0.548	0.891
FC6	0.635	0.832
FC7	0.861	0.561
FC6+FC7	0.921	0.881
POOL5+FC6	0.945	0.971
POOL5+FC7	0.941	0.972
POOL5+FC6+FC7	0.942	0.971

由于 SDL 高清航拍数据集的样本数量和采集场景不够丰富，我们没有能够测试检测器在各种不同场景中的虚警率。为了验证检测器的泛化能力，我们应该采集大量的航拍图像，在这些图像上进一步测试算法的性能。

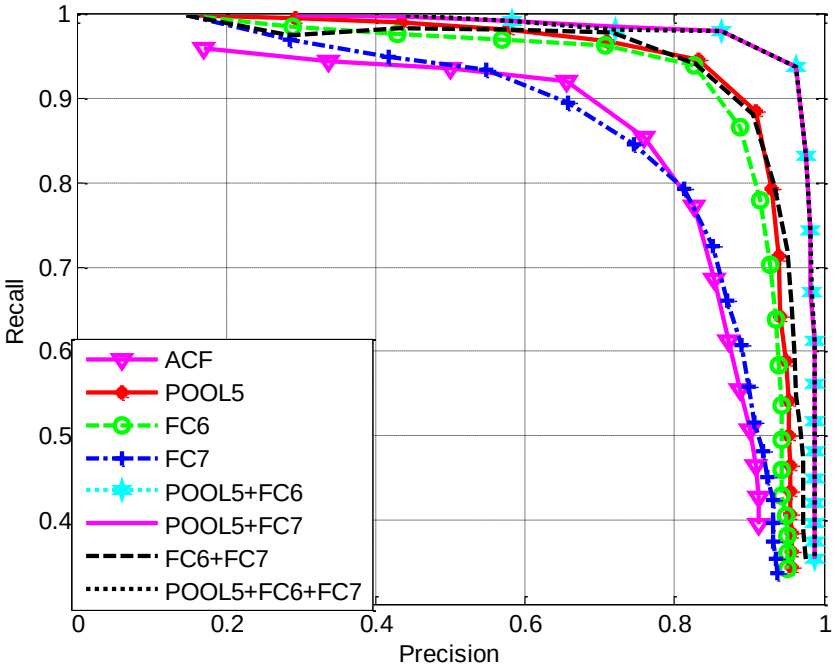

图 3-12 飞机检测性能曲线图

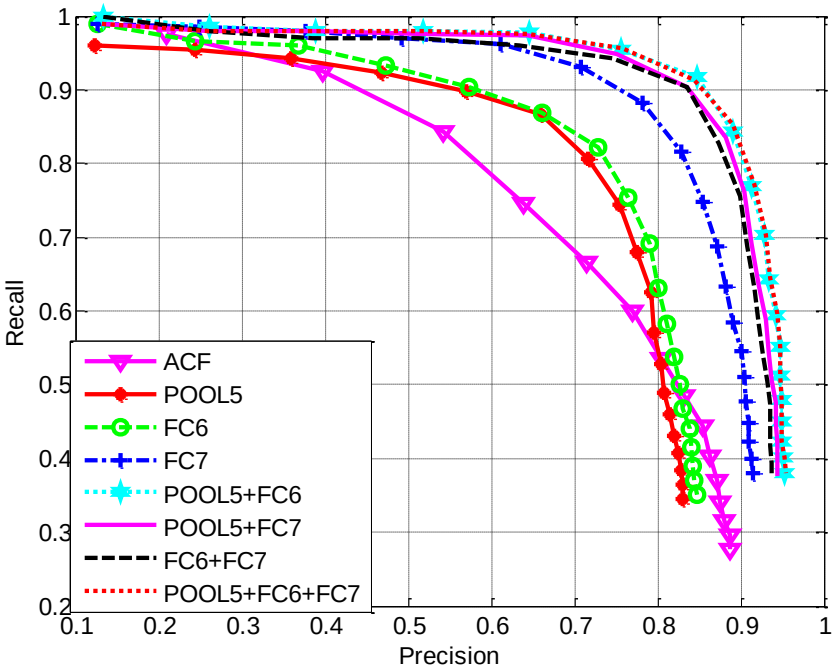

图 3-13 汽车检测性能曲线图

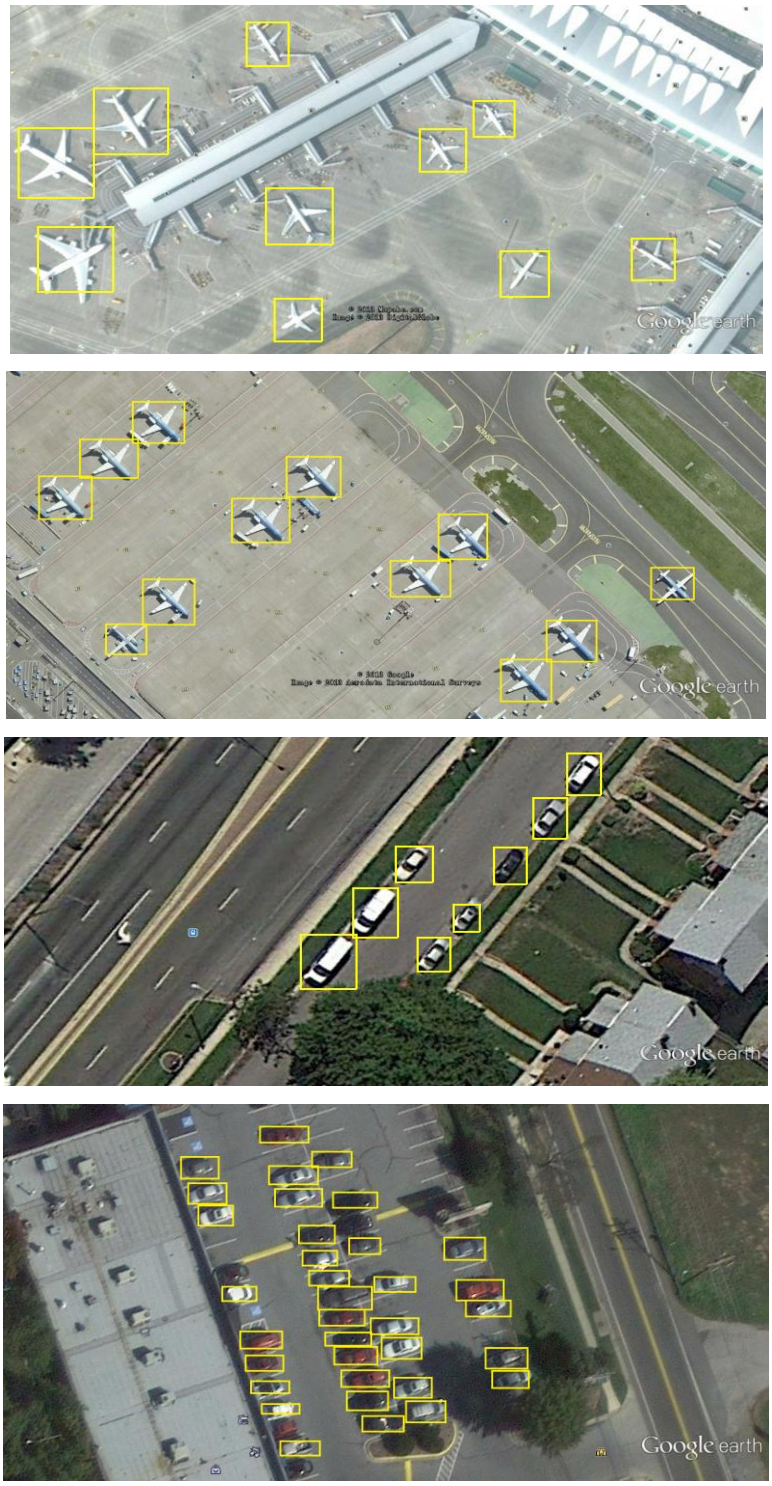

图 3-14 航拍目标检测结果示例

3.4 本章小结

针对航拍问题的难点，本章展开了一些研究。本章首先利用 Caltech 标定工具，标定了 SDL 高清航拍数据集。主要标定了飞机和汽车这两种常见的目标。基于标定的数据集，我们进行了一些尝试。我们采用 t-SNE 高维数据可视化工具分析了卷积网络特征的角度敏感性，基于对特征的的分析，选择了合理的混合卷积特征实现角度不敏感特征描述，并采用实验进行了验证。由于我们的特征能够对角度不敏感，因此我们考虑不再旋转原图进行检测，本文提出了非旋转原图的检测方案，即采用基于图像分割的预定位算法，再进行精确检测，算法成功实现了高清航拍图像中飞机和汽车的高效率检测，并取得了较好的检测性能。

第四章　弱监督高清航拍目标检测算法研究

在第三章中，实验证明，卷积神经网络特征可以作为一种角度不敏感特征来使用，基于深层卷积特征和大量的标定好的样本，能够训练出性能良好的检测器，对高清航拍图像中角度变化范围较大的目标进行准确的检测。

然而，第三章中的方法基于手工标定的数据集。需要标定数百幅含有汽车和飞机的航拍图像，大约 10 个学生花了 5 周的时间才标定完毕，耗费的人力是相当可观的。本章尝试基于多示例学习算法的弱监督学习算法。从弱标定的样本中挖掘样本的内在模式，并进行正例挖掘，让计算机帮助我们挑选正例样本图像，再基于样本图像训练检测器。

弱标定是指把包含有正例样本的图像标为正例，而不必精确到图像中目标的位置。而普通标定则需要标定每个飞机所在的位置。本文设计了一种基于弱标定数据的弱监督学习算法，基于弱标定的数据以及一些目标的模板图像（本文称为种子点），挖掘出大部分标定数据中目标的位置，作为训练样本使用，本文称这个过程为正样本挖掘。

4.1 弱监督航拍目标检测算法

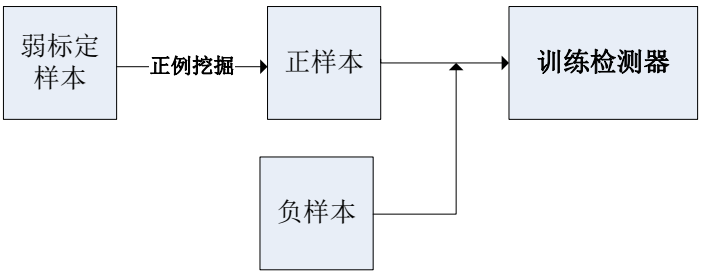

图 4-1 弱监督目标检测

如图 4-2 所示，传统的目标检测需要进行手工正样本标定。在标定样本的基础上，进行特征提取与分类器学习。在这个过程中，手工标定耗费人力较多，尤其是针对姿态多变的航拍目标，通常需要标定数千个目标才能训练出泛化能力较强的检测器。而弱监督高清航拍目标检测算法减轻了手动标定的负担。如

图 4-1，弱监督高清航拍目标检测算法以正例挖掘取代了手工标定，只要正例挖掘的准确率足够高，则检测性能就可以达到监督式目标检测的标准。算法 2 以及图 4-3 介绍了正例挖掘的过程：

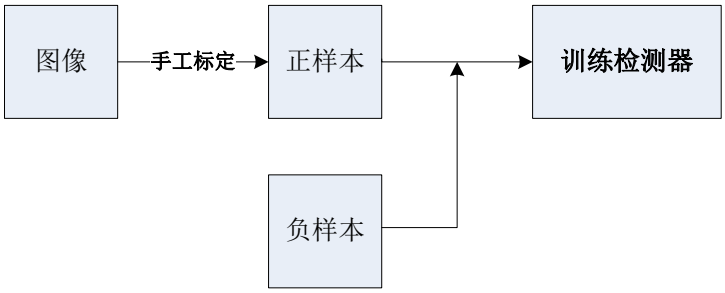

图 4-2 传统目标检测

算法 2 正例挖掘算法

输入：种子图像，K 幅弱标定的正例图像，反例图像

输出：K 个正包

1.用 Selective Search 算法对所有 K 幅正例图像和 N 幅反例图像求待选区域。

2.计算正例图像和反例图像的待选区域特征；计算种子点特征。

3.如果是第一次执行该段代码，那么分别计算种子点和 K 幅图像中图像块特征的距离，将每幅图像块和种子点最近的几个图像块放入相应的 K 个正包中（即每个正包中存放着该幅图像中的疑似目标）；如果不是第一次执行改行，那么用 4 中的 MISVM（Multi-instance Learning SVM）[25]分类器给 K 幅图像中的特征得分，删除正包中数据，将 SVM 得分最高的放入包中。

4.如果这一行执行次数小于 5，那么用 K 个正包中的图像块，反例的特征训练 MISVM[25]分类器；否则结束程序。

在正例挖掘的过程中，用到了 MISVM，即多示例支持向量机，它是一种分类器学习算法，其基本目标是从弱标定的数据中训练出一个能够识别正样本的分类器。

MISVM 是基于多示例学习的分类器学习算法。多示例学习[35,36,37,38]在包的

粒度对样本进行标记，每个包含有若干个示例（即正例样本）以及反例样本，这里的“包”即为弱标定样本。若某个包被标记为正包，则该包中至少有一个正示例；反之，若某个包被标记为负包，则该包中的所有示例为负示例。多示例学习的目的就是通过对这些包所包含的信息的学习与挖掘，尽可能准确地判断未知包的类别，MISVM 通过弱标定数据训练分类器，能够判断样本是正例的概率大小。

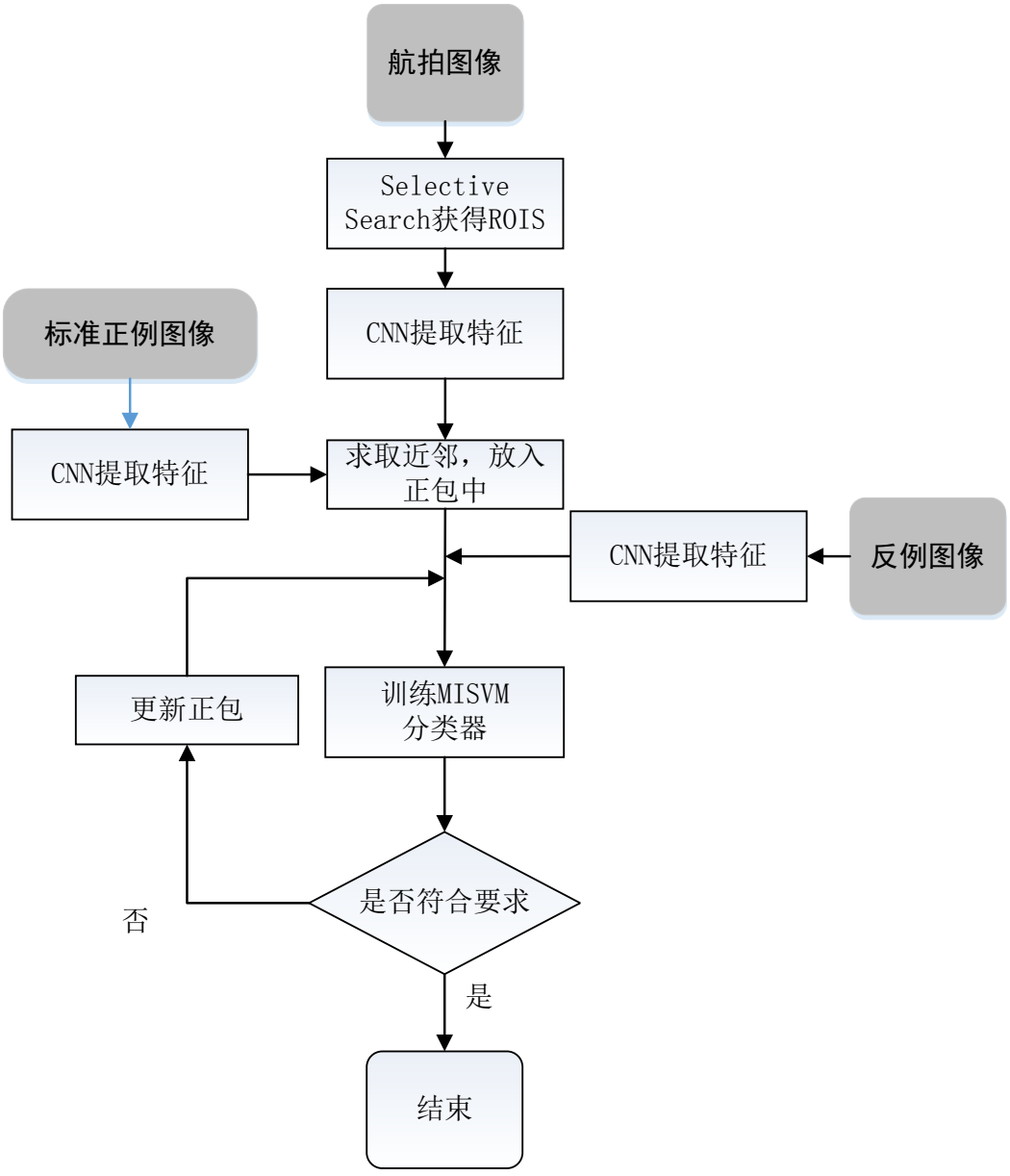

图 4-3　正例挖掘流程图

本文采用 MISVM 作为正例挖掘的工具，图 4-4 介绍了 MISVM 的详细流程。

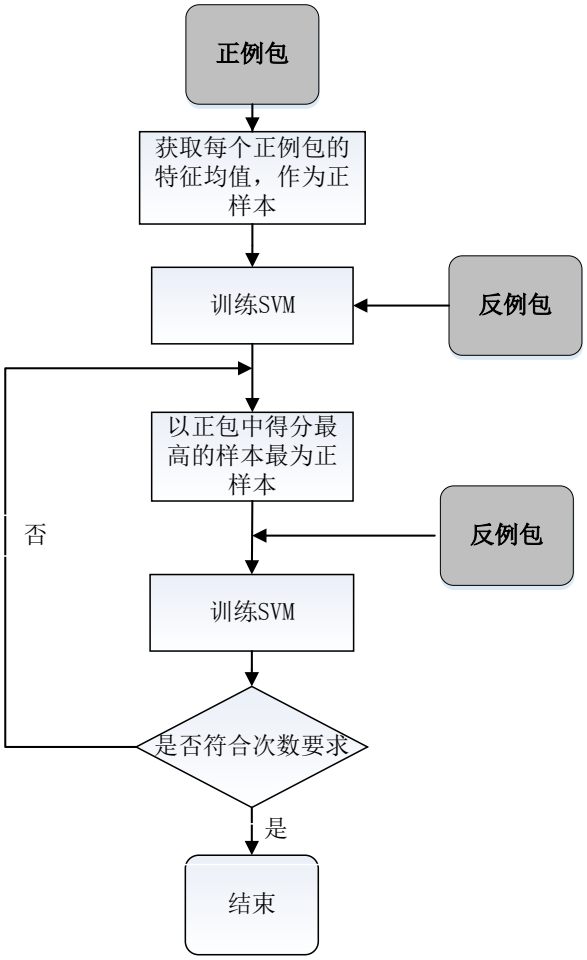

图 4-4 MISVM 流程图

在统计模式分类问题中，前提通常是给定了一些样本对，$(\mathrm{x}_i,\mathrm{y}_i)\in\mathbb{R}^d\times\mathrm{Y}$。而模式分类的目的是从给定样本的分布中获得规律，能够从模式映射到样本的类别。 即学得函数： $f:\mathbb{R}^d\to\mathrm{Y}$。目标检测问题中， $\mathrm{Y}=\{-1,1\}$。弱监督学习的区别在于放宽了对样本对的要求，样本对被放入一些包中，并且每个包有一个标签，而每个样本则不带有标签。数学上的描述如下：给定信息为：(B_i,Y_i)，其中 $B_I=\{\mathrm{x}_i:\mathrm{i}\in\mathrm{I}\}$， $I\subseteq\{1,2,...\mathrm{n}\}$，如果包 B_i 的标签 $Y_i=1$，那么包 B_i 的样本中最少有一个为正样本。而若 $Y_i=-1$，则 B_i 中全部为负样本。

求解模式分类问题的比较好的方法之一是 SVM，弱监督学习将 SVM 进行了

扩展。将边界最大化问题进行了推广，提出了多示例支持向量机（MISVM）。MISVM 将函数间隔推广为：

$$v_I = Y_I \max_{i \in I}(\langle w, x_i \rangle + \mathrm{b}) \tag{4-1}$$

这个公式认为，包的标签的预测值为：$\hat{Y}_I = \mathrm{sgn}\max_{i \in I}(\langle w, x_i \rangle + b)$，注意，正包的边界由 SVM 得分最高的样本决定，负包的边界则相反。基于包间隔的定义，SVM 的问题可重新定义为：

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_I \zeta_I \\ &s.t.\ \ \forall I : Y_I \max_{i \in I}(\langle w, x_i \rangle + \mathrm{b}) \geq 1 - \zeta_I, \zeta_I \geq 0 \end{aligned} \tag{4-2}$$

这个问题可以认为是一个混合整数规划的问题，优化过程如 4-4 所示，采用迭代循环的过程进行逐步优化。

由流程图 4-3 所示，深色的框是所需要的信息，可见，弱监督学习算法仅需要几幅正例样本作为种子点，对正包信息和反包信息进行反例挖掘，就可以输出正例样本。正包和反包的获取是比较容易的。

4.2 实验过程与结果分析

本文在 SDL 高清航拍飞机数据集中的飞机数据集中进行了实验；采集了{正包：50 幅，负包：50 幅，种子点 10 个}作为训练数据。

如图 4-7 可见，在挖掘的过程中，挖掘结果的准确率随着迭代次数的增加在逐步的提升。正包的初始准确率只有 45%左右，经过五轮的挖掘，正样本的纯净度已经达到 95%以上，已经基本可以作为监督学习的正样本使用。图 4-5 为第一轮挖掘后结果，图 4-6 位第五轮挖掘后的结果，可见，一些样本经过 5 轮的迭代后由错误变为正确；另外，第五轮的结果已经非常纯净，但是依然有一些飞机挖的不是很准确，如飞机的机尾容易被误认为是整个飞机。这可能是由于一些机尾和飞机的整体比较相似，导致学习器误判。这个问题可以通过对挖掘结果进行二次筛选进行精确挖掘，进而剔除挖掘错误的样本。

图 4-5 正例挖掘第一轮结果

图 4-6 正例挖掘第五轮结果

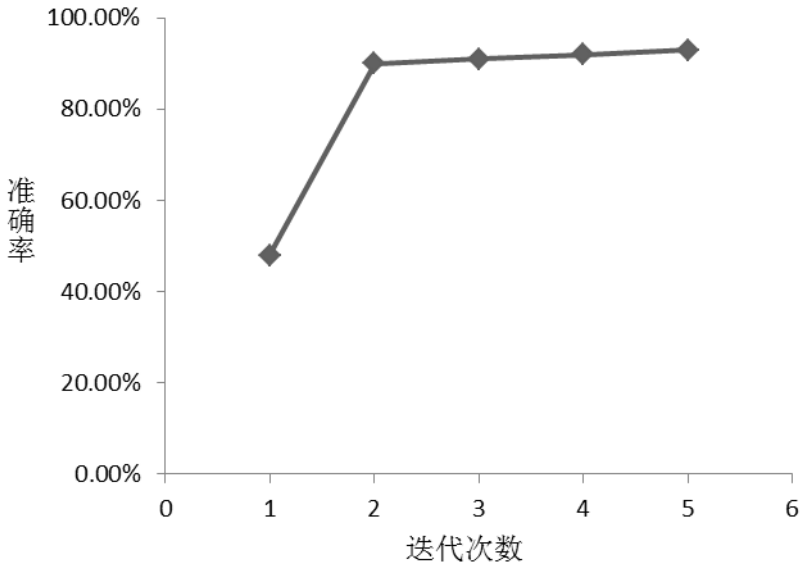

图 4-7 正例挖掘准确率与迭代次数的关系

4.3 本章小结

本章介绍了本文的一个尝试，即借助多示例学习算法实现正例样本挖掘。本文通过流程图介绍了算法的具体细节，通过实验验证了该算法的可行性并给出了实验结果。

基于弱监督学习的正例挖掘算法利用的信息是几幅目标的标准样本图和弱标定的数据集，在这些信息的基础上，对数据集进行反复的挖掘，以达到正例筛选的作用，这避免了手工标定的大量的工作量。

虽然在航拍飞机数据集上的实验结果达到了预期，但是因为时间关系，没有能够做更多的实验。在未来的研究中，从理论上证明该算法的收敛性将非常有意义。

第五章　结论与展望

航拍目标检测作为目标检测的子课题，是计算机视觉领域的重要研究内容。航拍目标检测在军事目标智能识别，遥感影像解析以及民用航空等领域具备广阔的应用前景。但是，航拍图像中的目标受颜色，长宽比变化以及复杂背景以及旋转的影响很大。本文提出使用深度卷积特征的混合来进行角度鲁棒的航拍目标检测。探索了深度卷积特征不同层的角度敏感特性。并且提出了基于图像分割的感兴趣区域提取方案。解决了航拍汽车和航拍飞机的卷积特征表达问题，解决这两种航拍目标旋转变化大，难以进行高速度检测的难题。

本文首先简述了航拍目标检测的研究背景和意义，国内外研究状况和本文的研究内容，然后在第二章中，详细地介绍了目标检测常用一些特征，并介绍了感兴趣区域提取算法和分类器。第三章介绍了本文的一些工作，本文提出了使用高维数据可视化工具来描述目标的特征与角度的关系。利用这种方式，本文从卷积神经网络的特征分析着手，提出了利用 AlextNet 的 POOL5 层提取航拍中的旋转不便特征描述子。在此基础上，采用基于图像分割的感兴趣区域提取算法对目标进行粗定位，再用卷积特征进行特征描述，最后用支持向量机分类。这个框架比原本的旋转原图的航拍目标检测框架有了很大的改进。在第四章中，针对训练目标检测分类器需要标定大量训练样本的问题，本文提出采用弱监督学习的方式进行正例挖掘，并在 SDL 航拍飞机数据集上做了实验，实验证明，本文提出的算法具备较好的正例挖掘的性能。

本文在高清航拍图像的目标检测中取得一定的阶段性成果，但因为时间关系，依然存在不足。在特征描述方面，求取深度卷积特征的计算量还是很大的，本文只尝试了深层特征，而没有尝试浅层特征，另外，可以尝试轻量级的神经网络，并探索卷积网络的角度不敏感性的原因；在分类器的选择方面，本文并没有提出改进的技巧，只是选择了支持向量机分类器。另外，本文只对航拍中比较清晰的高分辨率图像进行了一定的尝试，而低分辨率的图像本文并没有去实验。在正例挖掘中，由于时间仓促，算法还比较粗糙，没有能够细致地去进

行理论分析和算法设计，这一部分工作实验室会继续进行下去。

将来的高清晰度航拍图像目标检测算法可以进一步研究的方向可能会包括以下两个方面：第一个是提取复杂度较低的鲁棒特征，深度卷积神经网络由于层数多，计算量大，因此，很难能实用化；另外，在感兴趣区域提取方面，还有很大的发展空间，目前的算法能把穷举的几十万个感兴趣区域降低到几千，然而，即使是几千的感兴趣区域，计算它们的特征也需要一定的计算资源的开销，因此，高查全率，低错误率，高速率的感兴趣区域提取算法会成为一个比较重要的研究方向。

参 考 文 献

[1] Papageorgiou C., Poggio T.. A trainable system for object detection[J]. *International Journal of Computer Vision*, 2000, 38(1): 15-33.

[2] Dalal N., Triggs B.. Histograms of oriented gradients for human detection[C]. *In: Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, 2005:886-893.

[3] Dalal N., Triggs B., Schmid C.. Human detection using oriented histograms of flow and appearance[C]. *In:Computer Vision–ECCV, Springer Berlin Heidelberg.* 2006: 428-441.

[4] Dalal N.. Finding people in images and videos[D]. *Doctoral dissertation, Institut National Polytechnique de Grenoble*, 2006.

[5] Wang X., Han T. X., Yan S.. An HOG-LBP human detector with partial occlusion handling[C]. *In:Computer Vision, IEEE 12th International Conference*,2009: 32-39.

[6] 邓乃扬, 田英杰, 数据挖掘中的新方法-支持向量机[M]. 北京:科学出版社, 2004.

[7] Schwartz W. R., Kembhavi A., Harwood D., Davis L. S. Human detection using partial least squares analysis[C]. *In:Computer vision, IEEE 12th international conference*, 2009:24-31.

[8] Cheng M.M., Zhang Z., Lin W.Y., et al. BING: Binarized normed gradients for objectness estimation at 300fps[C]. *In: Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, 2014: 3286-3293.

[9] Uijlings J.R., et al. Segmentation as selective search for object recognition[C]. *In:Computer Vision (ICCV), IEEE International Conference*, 2011:879-1886.

[10] Low D.G.. Object recognition from local scale-invariant features[C]. *In:Computer vision, The proceedings of the seventh IEEE international conference*, 1999:1150-1157.

[11] Lowe D. G.. Distinctive image features from scale-invariant keypoints[J]. *International journal of computer vision*, 2004, 60(2): 91-110.

[12] Bay H., Tuytelaars T., Van Gool L.. Surf: Speeded up robust features[J]. *Computer vision–ECCV, Springer Berlin Heidelberg*, *2006*:404-417.

[13] Ma Y., Chen X., Chen G.. Pedestrian detection and tracking using HOG and oriented-LBP features[M]. *In:Network and Parallel Computing, Springer Berlin Heidelberg*, 2011:176-184.

[14] Krizhevsky A., Sutskever I., Hinton G.E.. Imagenet classification with deep convolutional neural networks[C]. *In:Advances in neural information processing system*, 2015:1097-1105.

[15] Papageorgiou C., Poggio T.. A trainable system for object detection[J]. *International Journal of Computer Vision*,2000,38(1):15-33.

[16] Viola P., Jones, M.. Robust real-time object detection[J]. *International Journal of Computer Vision*, 2001,4:34-47.

[17] Goldstein, E.. Sensation and perception[M]. *Cengage Learning*, 2013.

[18] Schapire R. E.. A brief introduction to boosting[J]. *Ijcai*,1999:1401-1406.

[19] Schapire R. E., Freund Y., Bartlett P., Lee W.S.. Boosting the margin: A new explanation for the effectiveness of voting methods[J]. *Annals of statistics*,1998:1651-1686.

[20] Kembhavi A., Harwood D., Davis L. S. Vehicle detection using partial least squares[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 2011,33(6):1250-1265.

[21] Van der Maaten L., Hinton G.. Visualizing data using t-SNE[J]. *Journal of Machine Learning Research*, 2008:2579-2605.

[22] Viola P., Jones M.. Robust real-time object detection[J]. *International Journal of Computer Vision*, 2001,4:34-47.

[23] Girshick R., Donahue J., Darrell T., Malik J.. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *In:Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, 2014:580-587.

[24] Dollar P., Wojek C., Schiele B., et al. Pedestrian detection: An evaluation of the state of the art[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 2012, 34(4): 743-761.

[25] Andrews S., Tsochantaridis I., Hofmann T.. Support vector machines for multiple-instance learning[C]. *In: Advances in neural information processing systems*, 2002: 561-568.

[26] Choi J. Y, Yang Y. K.. Vehicle detection from aerial images using local shape information[M], *Advances in Image and Video Technology.,Springer Berlin Heidelberg*, 2009: 227-236.

[27] Zhao T., Nevatia R.. Car detection in low resolution aerial images[J]. *Image and Vision Computing*, 2003, 21(8): 693-703.

[28] Eikvil L., Aurdal L., Koren H.. Classification-based vehicle detection in high-resolution satellite images[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2009, 64(1): 65-72.

[29] Zheng H, Pan L, Li L. A morphological neural network approach for vehicle detection from high resolution satellite imagery[C]. *In: Neural Information Processing. Springer Berlin Heidelberg*, 2006: 99-106.

[30] Grabner H, Nguyen T.T, Gruber B, et al. On-line boosting-based car detection from aerial images[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2008, 63(3): 382-396.

[31] Reed S., Sohn K., Zhang Y., et al. Learning to disentangle factors of variation with manifold interaction[C], *In: Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014: 1431-1439.

[32] Kavukcuoglu K., Sermanet P., Boureau Y. L., et al. Learning convolutional feature hierarchies for visual recognition[C], *Advances in neural information processing systems*, 2010: 1090-1098.

[33] Hinton G., Osindero S., Teh Y. W.. A fast learning algorithm for deep belief nets[J]. *Neural computation*, 2006, 18(7): 1527-1554.

[34] Imagenet 官网，2015-5-9 http://www.image-net.org/

[35] Wang C., Huang K., Ren W., et al. Large-Scale Weakly Supervised Object Localization via Latent Category Learning[J]. *Image Processing, IEEE Transactions*, 2015, 24(4): 1371-1385.

[36] Song H. O., Girshick R., Jegelka S., et al. On learning to localize objects with minimal supervision[C], *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 2014: 1611-1619.

[37] Li Y. F., Zhou Z. H.. Towards making unlabeled data never hurt[J], *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015, 37(1): 175-188.

[38] Cinbis R. G., Verbeek J., Schmid C.. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning[J]. *arXiv preprint arXiv*, 2015, 1503.00949.

[39] Dollár P., Appel R., Belongie S., et al. Fast feature pyramids for object detection[J], *Pattern Analysis and Machine Intelligence, IEEE Transactions* , 2014, 36(8): 1532-1545.

[40] Reed S, Sohn K, Zhang Y, et al. Learning to disentangle factors of variation with manifold interaction[C], *In: Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014: 1431-1439.

[41] http://ufldl.stanford.edu/wiki/index.php/UFLDL

[42] Tuzel O., Porikli F., Meer P.. Pedestrian detection via classification on riemannian manifolds[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 2008, 30(10): 1713-1727.

[43] Zheng S., Sturgess P., Torr P. H. S.. Approximate structured output learning for constrained local models with application to real-time facial feature detection and tracking on low-power devices[C]. *In:Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshop*s, 2013: 1-8.

[44] Hare S., Saffari A., Torr P. H. S.. Efficient online structured output learning for keypoint-based object tracking[C]. *In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference*, 2012: 1894-1901.

[45] Papageorgiou C. P., Oren M., Poggio T.. A general framework for object detection[C]. *In: Computer vision, 1998. sixth international conference*, 1998: 555-562.

[46]Viola P., and Jones M. Rapid object detection using a boosted cascade of simple features[C].*In：IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001:I-511 - I-518.

[47] http://www.cs.toronto.edu/~hinton/

[48] Krizhevsky A., Sutskever I., Hinton G. E.. Imagenet classification with deep convolutional neural networks[C]. *In: Advances in neural information processing systems*, 2012: 1097-1105.

[49] 杨乐坤教授主页. http://yann.lecun.com/

个人简介以及文章发表

个人简介：

朱海港　男　汉族　中共党员

2012.09-2015.07 中国科学院大学 计算机应用技术 硕士

2008.09-2012.07 西北工业大学　　电子科学与技术 学士

曾获荣誉：

•全国大学生数模竞赛二等奖　　（2010 年）

•全国大学生英语竞赛三等奖 （2010 年）

•西北工业大学一等奖学金　　（2009 年）

•西北工业大学校三好学生　　（2009 年）

•西工大电子爱好者协会优秀干部 （2011 年）

•苏北数学建模竞赛二等奖　　（2012 年）

已发表文章：

Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, Jianbin Jiao, “Orientation robust object detection in aerial images using convolutional neural network.”, IEEE Int'l Conf. Image Processing (ICIP), 2015.已接收

致 谢

在中国科学院大学攻读硕士期间，我在科研上经历了不少的挫折，也得到了很大的锻炼和收获。在毕业论文即将完成之际，我由衷地感谢给我帮助的老师和同学以及家人。

本课题的研究工作是在焦建彬教授、叶齐祥副教授以及韩振军副教授的悉心指导下完成的。感谢焦建彬教授在我攻读硕士学位期间从学习和生活各个方面给予的无微不至的关怀与指导，以及在我论文的撰写和修改中倾注的心血。感谢叶齐祥老师在科研的过程中对我的的指导和鼓励,感谢叶老师在我学习和科研遇到困难，难以前行的时候，鼓励我，并为我提供思路。感谢韩振军老师在论文写作的过程中给予的指导和帮助以及生活上给予的关心。恩师们在科研上精益求精，在学术上认真严谨，他们的科学精神令人敬佩；恩师们在生活上关心爱护学生，和蔼可亲、平易近人，他们春风化雨的教诲让人感动。

感谢陈孝罡师兄在我遇到困难时帮助我解决难题，感谢他带我走进深度学习算法的世界，带我在目标检测领域入门，帮助我解决困难，在我一筹莫展的时候指点迷津。感谢各位师兄、师姐，在我研究生的三年中给予的学习上的耐心引导与生活中的种种帮助。感谢我的同届好友以及师弟师妹们，三年的科研生活中大家相互帮助、献计献策、相互提点，一起渡过了三年快乐的日子。这些快乐的时光在我记忆中永远不会褪色。

感谢我的父母亲人以及女朋友，他们给了我巨大的、无私的爱和永远无条件的支持，永远是我最坚强的后盾和避风的港湾，愿他们永远平安健康。

感谢参加开题及中期评阅的各位老师和专家们，他们丰富的经验和无私的工作对论文方向和研究进度的把握和指点给整个研究工作带来了许多帮助。

朱海港

2015.5