密级:_____

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

硕士学位论文

基于稀疏重构与卷积特征的人体目标检测

作者姓名：张耀

指导教师：叶齐祥副教授 中国科学院大学电子电气与通信工程学院

学位类别：工程硕士

学科专业：工业工程

研 究 所：中国科学院大学工程管理与信息技术学院

二〇一五 年 五 月

Pedestrian Detection Based on Sparse Reconstruction and Convolution Feature

By

Yao Zhang

A Thesis Submitted to
The University of Chinese Academy of Sciences
In Partial Fulfillment of the Requirement
For the Degree of
Master of Industrial Engineering

College of Engineering and Information Technology
University of Chinese Academy of Sciences
May, 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

随着计算机视觉基础理论的推广与计算机软硬件水平的快速发展，对目标进行动态检测在智能交通系统、智能监控系统、军事目标检测以及人工智能机器人等方面具有广泛的应用价值。人体检测是目标检测中的重要组成部分，是目标检测中比较典型的研究问题之一。由于人体检测受到姿态多变性、多视角以及遮挡光照变化等问题的困扰，检测性能仍然较低，难以实际应用。

视觉目标检测是多类分类问题，其步骤主要分为提取特征与训练分类器两部分。在特征提取方法中，主要分为手工设计的特征和基于学习的特征两类，每一类特征都有其各自的优势与不足。本文对人体检测系统中关键的特征提取部分进行了探索，并在所采集的高分辨率视频数据上进行实验以对提高数据分辨率是否会提高检测效率进行验证。

本文的主要工作如下：

（1） 研究了常用的手工设计特征以及基于学习的特征，根据其特征提取方法和应用，分析各类特征的优缺点。

（2） 基于稀疏编码算法与主成分分析法，结合多通道下的图像表征，设计将两类特征结合生成新的特征描述子进行人体目标检测，分析结合两类特征的可行性与互补性，同时在实验中研究了卷积正交的滤波模板对特征表达的影响。

（3） 采集高分辨率图像视频数据集，利用现有的特征进行实验，分析提高数据集分辨率对于提高检测准确率的影响。

实验结果表明，结合两类特征具有一定的可行性。考虑到两者之间的互补性，本文提出的特征描述子，在 INRIA 和 Caltech 数据集上的误检率分别比 baseline 降低 3%与 13%。同时，通过在高分辨率图像数据上实验发现，在提取更加有效的特征描述子前提下，提高数据分辨率能够进一步提高检测效率。

关键字：人体检测，特征提取，稀疏重构，卷积特征，高分辨率数据

Abstract

With the promotion of the basic theories of computer vision and the rapid development of computer software and hardware, object detection is of much value in intelligent transportation system, intelligent monitoring system, military object detection, robot of artificial intelligence and so on. Pedestrian detection is an important part of object detection, and is one of the typical research questions. However, due to the interference of multi-pose, multi-view and occlusion of pedestrian, pedestrian detection is difficult to apply.

Object detection is a classification problem. Feature extraction and training of the classifier are the two steps of it. According to the methods of feature extraction, there are mainly two classes. One is hand craft feature, and the other is learning feature. Each feature has its own advantages and disadvantages. In this thesis, we carried out the exploration to the feature extraction, and conducted experiments on the high resolution video data to analyze whether high resolution is helpful to pedestrian detection.

In this thesis, the main work are as follows:

(1) We studied common hand craft features and learning features. According to the extraction methods and application, we analyzed the advantages and disadvantages of two kinds of features.

(2) Based on sparse coding algorithm and principal component analysis, we combined with multi-channel image feature to design new feature descriptors of pedestrian detection. In addition, we analyzed the feasibility of combining two kinds of featrue, and studied the influence of convolute orthogonal filter templates on the feature.

(3) We collected high resolution video dataset, and conducted experiments using existing features to analyze the influence of using high resolution dataset to improve the accuracy of detection.

The results of experiments show that the combination of two kinds of features has certain feasibility. Considering the complementarity and combining features to generate description, we reduce missrate by 3% and 13% in INRIA and Caltech dataset

respectively compared with our baseline. At the same time, through the use of high resolution image data for experiments, we found that under the more effective feature, improving data resolution will improve the efficiency of detection.

Key Word: Pedestrian Detection, Feature Extraction, Sparse Reconstruction, Convolution Feature, High Resolution Video Dataset

目录

摘　要 I
Abstract III
目录 V
图目录 VII
表目录 IX
第一章　绪论 1
1.1　课题背景与研究意义 1
1.2　国内外研究现状 3
1.2.1　手工设计的特征 4
1.2.2　基于学习的特征 6
1.3　本文研究内容 7
1.4　本文的组织结构 8
第二章　人体目标特征提取综述 9
2.1　特征选择 9
2.2　手工设计的特征 10
2.2.1　Haar-like 特征 10
2.2.2　HOG 特征 13
2.2.3　LBP 特征 14
2.2.4　SIFT 特征 16
2.3　学习的特征 19
2.3.1　卷积神经网模型 19
2.3.2　稀疏特征选择 21
2.4 本章小结 23
第三章　基于稀疏重构与卷积的人体检测研究 25
3.1　整体研究框架 25
3.2　基于集合的多通道特征表达检测方法 26
3.2.1　通道的选取与表达 27
3.2.2　分类器的选择 28
3.3　基于多通道的稀疏重构特征表达方法 30
3.3.1　稀疏重构的特征表达 30
3.3.2　实验方法 31
3.3.3　实验结果与分析 34
3.4　基于多通道的卷积正交模板特征表达方法 38
3.4.1　主成分分析方法 38
3.4.2　实验方法 40

3.4.3 实验结果与分析 42
3.5 本章小结 45
第四章 高分辨率数据在人体检测中的研究 47
4.1 问题分析 47
4.2 人体检测数据库 48
4.3 实验方法与对比结果分析 51
总结与展望 55
参考文献 57
个人简历及论文发表 61
致 谢 63

图目录

图 1-1　人体检测的一些典型应用 2
图 1-2　四幅图像分别对应四个难点 3
图 2-1　目标检测框架 9
图 2-2　Haar-like 特征 11
图 2-3　计算积分图 12
图 2-4　拓展的 Haar-like 特征 12
图 2-5　原始图像与 LBP 图谱 15
图 2-6　DoG 计算图 17
图 2-7　特征点周围的窗口分解，并为每个子窗口创建 8 位直方图 19
图 2-8　CNN 的结构示意图。C_1 与 C_2 是卷积层，P_1 与 P_2 是采样层 20
图 2-9　深度卷积神经网络 21
图 3-1　实验整体框架图 26
图 3-2　ACF 实验整体框架 27
图 3-3　人体图像的多通道信息提取 28
图 3-4　决策树构成的 AdaBoost 分类器 30
图 3-5　手工设计特征方法下的多通道特征提取 32
图 3-6　实验流程图 32
图 3-7　不同通道下原样本（第一层）和重构后样本对比图（第二层） 33
图 3-8　编码特征在 INRIA 与 Caltech 上的性能曲线 34
图 3-9　利用编码作为特征的实验结果（左）与 ACF 实验结果（右） 36
图 3-10　重构特征在 INRIA 与 Caltech 上的性能曲线 37
图 3-11　实验结果图 37
图 3-12　常用特征在 INRIA 上的性能对比 38
图 3-13　十通道下的 PCA 滤波模板，每列一组包含 8 个特征向量 41
图 3-14　各通道下卷积滤波模板的示意图 42
图 3-15　卷积正交模板特征在 INRIA 与 Caltech 上的性能曲线 42
图 3-16　检测结果示意图 43
图 3-17　常用特征在 INRIA 上的性能曲线 44

图 4-1　INRIA 中训练集（左）与测试集（右）数据.....49
图 4-2　Caltech 中视频图像.....50
图 4-3　Pri-SDL 高分辨率人体数据集.....51
图 4-4　ACF 与重构特征在高分辨率数据集的性能曲线.....52
图 4-5　卷积正交模板特征在高分辨率数据集的性能曲线.....52
图 4-6　检测结果示意图.....53
图 4-7　漏检图样.....54

表目录

表 3-1　参数对比结果.....44
表 4-1　人体检测公开数据集表.....48

第一章　绪论

1.1　课题背景与研究意义

计算机视觉的研究为了是让机器通过外设收集获得周围环境中物体目标的信息，例如形态、位置、存在方式以及所处状态等，存储这些信息并进行理解与分析。图像与视频中的目标检测是指计算机或相关设备通过人为意志的设定对特定对象进行检测的活动,是计算机视觉中一个非常重要的研究领域。科技的发展与现实需求的提高，使得人们获取的信息量在急速增加，如何快速有效地让人们从大量繁杂的信息中找到关键内容成为了一种迫切的需要。随着计算机视觉基础理论的推广与计算机软硬件水平的快速发展，对目标进行检测并实时跟踪逐渐成为研究的热点，其在智能视频监控、智能交通辅助、医学导航、军事目标检测以及人工智能机器人等方面具有广泛的应用。

人体检测是目标检测领域的重要组成部分，是比较典型的研究问题之一。而人体检测又不是一个孤立的问题，它通常与行人路径跟踪、人体行为分析以及集群场景理解等结合，因此具有很高的理论意义和应用价值。一些典型的应用包括：

（1） 智能视频监控

近年来，随着社会的快速发展与国家基础设施建设的完善，各行各业的安全防范意识逐步增强，重点行业对于安防与报警系统的需求更是与日俱增，因此视频监控在现实生活中应用非常广泛。虽然视频监控系统已经广泛应用于人群较多的公共场所，但是依旧需要大量的人为操作才能完成监控任务，没有充分发挥监控系统的智能性和主动性。尤其是在人流量较大的地方，很难及时发现各种异常情况。因此利用人体检测技术，逐步实现视频监控的智能化，不仅可以弥补人力的不足，还可以提高监控准确率，减少损失。

（2） 智能交通辅助

随着经济的发展，现代社会的车辆总数在急剧增加，而人们对车辆驾驶的更高要求和随之而来的交通隐患使得在车辆控制系统与交通监控系统中需要更加

高效安全的方式来保障。当路面发生交通事故，如何快速发现事故位置，如何快速使得交通恢复畅通，这时就需要交通系统高效快速的发现并定位受伤的人群与车辆，及时通知工作人员进行处理，尽量降低人群与财产的损失，同时尽早恢复交通。

（3） 智能人机交互

在电子设备上实现人机交互的基础模型已经成为现实。当前计算机技术的进步，使得语音识别、汉字识别等功能在快速发展，对身体动作的识别、周围环境的感知以及手势和表情的感知能进一步提高人机交互的性能。

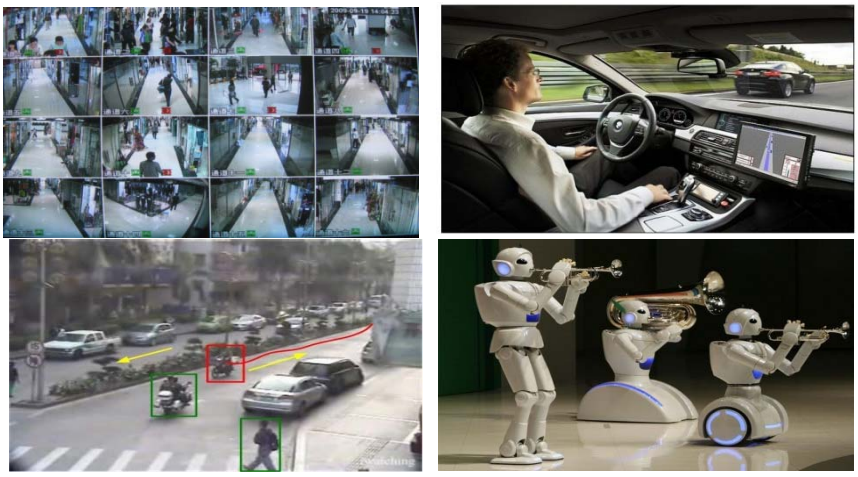

图 1-1　人体检测的一些典型应用

对于不同的应用背景，人体检测的研究各有自身的特点。对于室外目标，容易受到自然环境和活动范围的干扰；在集群场景中，容易与其他的目标产生遮挡。Dalal 和 Triggs[1]于 2005 年提出基于梯度直方图（Histogram of Orientated Gradient，HOG）特征与支撑向量机（Support Vector Machine，SVM）结合的行人检测算法，使得人体检测性能有了较大提升。X. Wang[2]等人于 2009 年结合梯度直方图与局部二值模型（Local Binary Pattern，LBP）提出了一种新的组合特征 HOG_LBP 特征描述子，并且提出了一种解决部分遮挡难点的方法，使得人体检测性能又有了较大的提升。P. F. Felzenszwalb[3]等人在 2010 年提出的部分可形变的模型（Deformable Parts Model，DPM），解决了特征对齐和模式分散的问题。

虽然人体检测理论研究已经日渐成熟，但是在存在部分遮挡以及低分辨率、人体多姿态等情况下，人体检测性能仍然较低，难以应用到实际环境中[4]。因此，在进行人体检测中将多信息进行融合，从而提高人体检测的鲁棒性、准确性，具有重要的理论意义与应用价值。

本文受到了以下课题的资助：

（1）“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，（课题编号：61039003）

（2）“多视角多姿态人体目标检测”，国家自然科学基金委面上项目，（课题编号：61271433）

1.2　国内外研究现状

人体检测的方法研究早在上世纪 90 年代就开始了。在过去十几年中，人体检测的发展呈现出以下趋势：1) 训练样本数据量急剧增加，样本内容复杂多样；2) 检测性能不断提高，由最初的每帧几秒到每秒几十至上百帧，检测的效率和精度也有很大提高；3) 传输图像视频的设备由单一化向多类化转变，特征提取的发展趋势由单一特征转向特征融合。

由于人体检测的问题与计算机视觉很多方面都有联系，如特征提取、分类器训练等，而且要面临很多现实环境带来的困难，因此，在检测过程中遇到了一些难题，主要包括以下几点：

(a)　(b)　(c)　(d)

图 1-2　四幅图像分别对应四个难点

（1） 人体的姿态是多变的，每一个部分都可能处于运动状态，很难找到一个固定的方法进行建模。除此之外，拍摄方向也会引出人体姿态的多样性。

（2） 人体外观具有多样性。高矮胖瘦、着装肤色、光照背景的变化，使得在纹理方面没有统一的模式，因此也加大了人体检测难度。

（3） 含有人体的图像，背景复杂且多样化。同一目标人体所处环境的背景不同，在特征提取中会形成不同的特征表达，因此产生严重干扰。同时低分辨率的视频图片也会影响检测的结果。

（4） 现实生活中，人体目标很容易被其他人体或者物体遮挡，如房屋、车辆等，这些干扰也会产生误检和漏检，降低检测性能。

面对人体检测中出现的诸多问题，逐步出现了多种人体目标检测的研究方法。首先是基于背景模型人体检测方法。该方法优势在于操作简单，将目标图像与背景图像进行对比，将不同的部分标出即可，但是很容易受到外界环境影响，在复杂环境下很难实现好的检测效果。其次是基于人体模型匹配的检测方法，主要是通过建立人体模型，与目标图像各个区域进行匹配，寻找人体目标，但该模型计算复杂，效率低下，在繁杂背景下很难应用。第三种是基于统计学习分类的人体检测方法，主要包含两个过程，即特征提取与分类器训练。首先通过对样本中目标人体进行特征提取，采用模型方法训练样本数据进而学习建立分类器，然后利用分类器对输入图像进行人体检测。该方法具有较好的鲁棒性，由于其较高的速度和准确率，已经逐步成为人体检测的主流。

综合近几年的发展，可以将提取特征的方法归为以下两类：手工设计的特征（hand craft feature）和基于学习的特征（learning feature），下面分别叙述。

1.2.1 手工设计的特征

手工设计的特征是按照人为设定的特征计算方法，对样本进行指定操作运算从而得到的表示方式。因其设计简单、检测速度快而备受关注。最初由 Papageorgiou 和 Poggio[5]等人提出将 Haar 小波函数作用于训练样本，从而得到基于灰度差的 Haar-like 特征。Viola[6]在积分图上计算 Haar 特征，利用 AdaBoost

进行分类，达到了实时检测的目的。随后许多研究者逐步参与到该特征的改进中，用完备的 Haar-like 特征分别用来表示人脸与人体特征。虽然 Haar-like 特征对于人脸表示的效果比较好，但它不太适合表示人体目标，原因在于 Haar-like 特征比较适合描述目标的显著区域，如眼睛、嘴巴、眉毛等，不太适合表示边缘轮廓信息，因为其容易受到目标形态，光照条件及视角的影响。2005 年，Dalal[1]提出了稠密的、重叠的、固定尺度的 HOG 局部特征描述子表述人体。该描述子借鉴了旋转尺度不变特征[7]（Scale Invariant Feature Transform，SIFT）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[8]。改进的 HOG 特征认为原始的 HOG 由固定尺度、固定位置的特征块组成，存在不能很好地把握人体局部轮廓特性的弊端。改进的 HOG 特征在人体检测中获得了不错的结果。O. Tuzel[9]等人使用区域的协方差算子（COV）表示人体特征。区域中的像素是由灰度、梯度等信息组成的特征向量，每个区域的 COV 算子是由位于该区域的所有像素点特征向量构成的协方差矩阵，协方差矩阵可以把握不同位置、不同尺度下人体区域的表征。Mu[10]等人认为原始的用来表示纹理特征的 LBP 算子，虽然在人脸识别、纹理检测等方面取得很好的效果，但不适合于描述人体的轮廓，因此提出使用改进的 LBP 算子来描述人体。X. Wang[2]等人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，对单调的灰度变化有不变特性；当背景比较复杂，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 特征融合表示人体目标，可以取得较好的检测结果。William[11]等人用基于边缘、纹理、颜色三种信息组合的高维描述子来表示人体模式。Wu[12]等人提出 Edgelet 特征表示人体模式，每个 Edgelet 特征是一条边，反映着人体局部位置的轮廓细节信息。Dollar 基于前人的研究，结合积分图和 HOG 提出了集合的多通道特征（Aggregate Channel Features，ACF）[13]，利用级联决策树构建 AdaBoost 分类器，进一步提高了人体检测精度。Nam[14]等发现级联决策树与没有相关性的特征配合更密切，对 ACF 每个通道进行局部

去相关，使得 ACF 在行人检测性能上有很大提高。除此之外，Tuzel[9]用局部块的协方差矩阵对目标进行描述，并将其在黎曼流形上进行分类。

1.2.2 基于学习的特征

随着稀疏表示和深度学习的发展，通过学习的方式进行目标表示也越来越受研究者们青睐。基于学习的特征是通过使用学习模型，对样本图像进行学习以及信息反馈，最终得到一种最优的表达方式。基于学习的特征是近年来出现的一种新方法，但迅猛的发展速度已使其成为一种主流特征提取方式。最具代表性的工作是 Girshick[15]将卷积神经网（Convolutional Neural Networks，CNN）引入目标描述，在候选区域上提取 CNN 特征并用 SVM 进行分类。该方法[15]在多目标检测数据集 PASCAL 上的分类性能超过同期的任何方法。CNN[16,17,18]是通过反向传播网络自主地学习所有的节点及参数，而一些非反向传播学习网络也取得了不错的成绩。Sermanet[19]等人通过字典学习获得卷积模板，在构建双层卷积网络的同时提取全局和局部特征。Ren[20]等人提出通过字典学习对局部块的重构系数进行直方图统计构建稀疏编码直方图（Histogram of Sparse Coding，HSC），完全采用和 HOG 一样的检测框架，但在行人检测上性能比 HOG 提升很多。Lim[21]等人用非监督学习的方法获得精确描述目标物体轮廓的中层特征，该方法在训练时依赖于人工标定的轮廓。Zhang[22]等人通过在平均梯度人体上提取 Informed Haar-like 模板对 ACF 进行卷积，和 ACF 一样采用级联决策树构建的 AdaBoost 分类器第一次将行人数据集 Caltech 上的高于 50 像素的行人检测误检率降到 35%以下。遗憾的是这种方法扩展性不好，对于新的目标物体需要重新学习滤波模板。

2012 年以来，随着大规模图像分类识别数据集合 ImageNet 的发布，深度卷积神经网络（Deep Convolutional Neural Networks，DCNN）的应用被推向一个新高潮。DCNN 网络提取的特征在其他领域中体现出了较好的性能。如基于区域预定位和 DCNN 特征分类的多类目标检测就是一个非常成功的例子。RCNN[15]（region with CNN feature，RCNN）利用 ImageNet 中的强大的深度卷积神经网络作为提取特征的工具，利用当时最好的 Selective Search[23]作为预定位算法，并

结合 SVM 分类器，刷新了目标检测的性能记录。

人体目标表示方法，尤其是目标表示的方法在近十几年来层出不穷。计算简单的方法对人体目标的表示能力往往不足，而表示能力好的方法往往计算复杂度很高。在希望不断提高检测性能的背景下，基于学习的特征虽然复杂[24,25,26]，但依然成为了人体目标表示的一个趋势。

1.3　本文研究内容

当前的人体检测系统通常是先提取感兴趣区域，然后提取特征并训练分类器进行人体检测，评测采用的数据集是传统的分辨率较低的 INRIA 数据集和 Caltech 数据集。在本文中，对人体检测系统中关键的提取特征部分进行了探索。除此之外，由于随着硬件设备性能的提升，视频图像的分辨率也逐步提高，因此本文采集了新的高分辨率数据并在此数据集上测试图像清晰度和提高检测准确率之间的关系。

本文的主要研究内容如下：

（1） 对常用的手工设计特征以及基于学习的特征，包括 HOG 特征、LBP 特征、CNN 特征等，进行了详细分析。根据其特征提取方法和应用，比较了两类特征的优缺点。

（2） 重点研究稀疏编码（Sparse Coding）生成的特征与 ACF 特征，并考虑两者的差异性与互补性，将两种特征有效融合形成新的描述子进行人体检测，实验结果表明可提高检测的准确率。

（3） 研究主成分分析法（Principal Component Analysis，PCA）等相关正交算法，并融合 ACF 特征，设计了多通道下的卷积特征，有效地提高了检测效率，分析了正交性在其中的作用。

（4） 与现有的低分辨率数据集形成对比，采集高分辨率视频数据，采用现有的特征进行实验，分析提高视频数据分辨率是否会有助于提高检测准确率。

1.4 本文的组织结构

第一章，绪论。介绍基于稀疏重构与卷积的人体目标检测研究背景和意义，分析了国内外人体检测的研究现状，尤其是分析了特征提取方面已有的相关研究，总结了人体检测尚未解决的一些问题，列出本文的主要研究内容。

第二章，人体目标特征提取综述。论述特征描述子的意义，分析人体检测中应用最普遍的两类特征的主要方法，主要包括特征的定义、实验方法和可以应用领域。

第三章，基于稀疏重构与卷积的人体检测研究。论述两类特征的优缺点以及结合的意义，提出结合多通道特征，利用稀疏重构方法与卷积正交滤波模板的方法生成新的特征描述子，采用结合后的特征进行实验，并展示其结果和性能对比图，分析特征结合的有效性与意义。同时，还研究正交算法在特征提取方面的结合应用，讨论其带来的优点与性能提升。

第四章，高分辨率数据在人体检测中的研究。采集高分辨率数据集，结合现有的特征方法进行实验，探索分析高分辨率数据是否会在提升检测性能方面带来帮助。

最后对本文的主要工作进行总结，分析了论文的不足、列举了仍然存在的难点问题，展望未来可以继续的研究方向。

第二章　人体目标特征提取综述

2.1　特征选择

视觉目标检测在模式识别中，就是二类或多类分类问题，通过特定的方式将检测目标与其他对象区分开。从本质上主要包含两个部分：目标表示与目标定位。目标表示要回答目标是什么，目标定位要回答目标在哪儿。目标检测的框架如图2-1所示。在训练阶段将待检测目标作为正例样本，背景作为反例样本，按照一定的特征提取方式将目标从图像空间映射到特征空间中，再利用分类器训练分类模型，将正例与反例样本分开。在检测阶段，从图像中获得候选窗口后，同样需要按照相同的特征提取方式将其映射到特征空间，利用训练好的模型进行分类检测，实现目标定位。由此可见，特征的选择与表达是决定分类器性能的重要因素，也是目标检测整体系统中的关键。

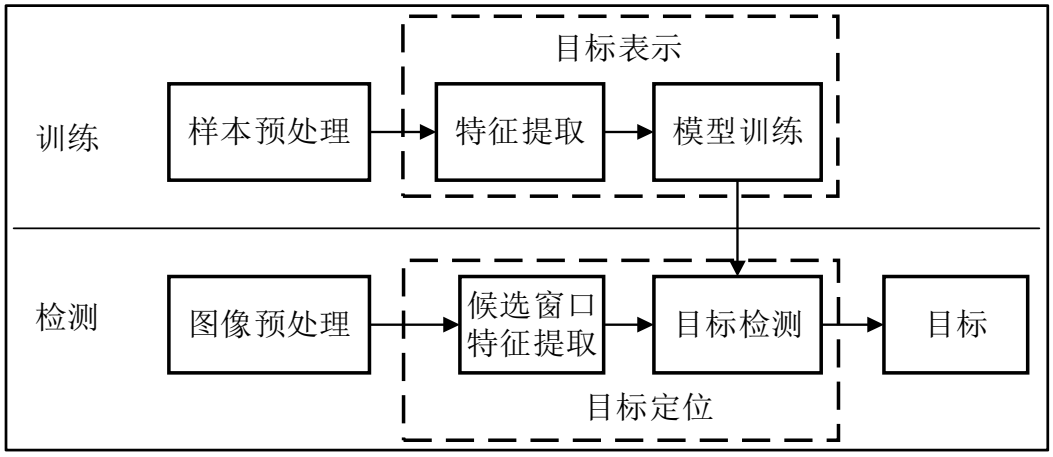

图 2-1　目标检测框架

在现实生活中，采集数据中的人体所在背景与环境在不断变化，还受到光照与气候的影响，人体服饰和姿态也不同，具有多特征、高噪声、非线性的特点。在检测中，我们需要根据不同场景和目标类型，选择更有代表性的特征，使得待检测目标与其他对象之间具有最大的区分性。选择主要分为两种：一种是从原始图像中选择一些具有代表性的点、边或者区域；另一种则通过变换的方式将原始

图像进行映射与组合，变换为新特征。在保持特征数量与维度尽量低的前提下，使得不同类别对象之间具有很强的区分性。

在早期的人体检测研究中，常用的提取底层特征方法有以下几种：

（1） 边缘检测

边缘通常是图像中亮度快速变化的部分，而这些地方能够有效的反映出图像中的信息。通过对图像进行卷积以及基于梯度的检测方法，能够得到图像中亮度变化较快的区域，对于图像分割与目标检测有很大帮助。

（2） 颜色区分

颜色是图像最基本的组成元素，对不同对象的识别，通常都是由其反射光线的性质来决定。不同的目标在相同场景下呈现的颜色是不同的。在数字图像处理中，常用 RGB 颜色空间来表示，除此之外，HSV、YUV 等颜色空间可以通过相互转换，根据不同场景，都可以用于对目标进行表示。

（3） 区域纹理

纹理主要指在图像一定范围内形成的具有规律性排列的图案。这种纹理能够一定程度上反映出图像自身的信息，增加目标的区分性。

此外还有光流、角点等可用于特征表达，并在特征应用中取得了一定效果。但是在人体检测中，单纯依靠上述这些简单特征方法已无法克服人体检测中的难点，同样也无法满足要求越来越高的检测效果。近年来，涌现出很多有效的新特征描述方法，从其表示方式主要分为两类，一类是手工设计的特征，如 HOG、LBP 等，另一类是通过模型自动学习得到目标特征，如卷积神经网模型、稀疏编码表示方法等，两类特征已逐步获得了很多较好的效果。

2.2 手工设计的特征

2.2.1 Haar-like 特征

Haar-like 特征是由 Haar 小波演变而来，首先是在表示人脸上获得实用，Viola[6]在其基础上，改进使用四种格式共三种类别的特征表达，分别是 2-矩形特

征、3-矩形特征和 4-矩形特征，如图 2-2 所示。Haar 特征模板以矩形为主，模板中只有黑色与白色，并规定这个矩形块的特征值即为两个矩形区域内像素和之差。通过 Haar 特征值的表达，可以得出灰度在图像中的变化情况。例如：人脸的一些区域可以用矩形叠加表达，眼睛的颜色相较额头的颜色深一些，而鼻子主体的颜色比其两侧的颜色要浅，嘴巴与周围皮肤相比颜色也要深等[27]。然而矩形特征相对简单，只能够描述少数结构，例如水平方向、垂直方向与对角方向，对边缘和线段表达性能较好，对于其他结构则不能很好的表达。

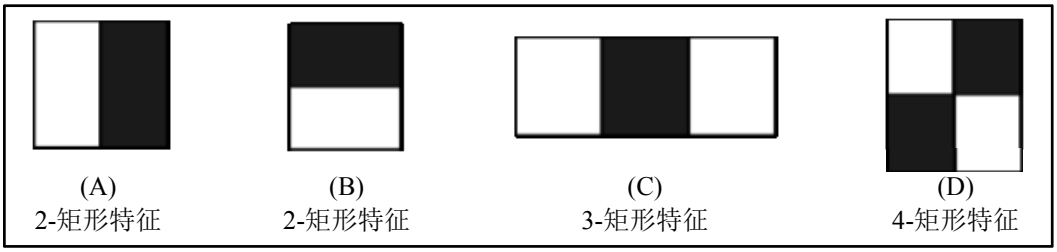

图 2-2　Haar-like 特征

图 2-2 中的 A，B 和 D 的特征值计算方法为：v=Sum 黑-Sum 白，C 的特征值计算方法则为：v=2*Sum 黑-Sum 白。Sum 黑表示矩形黑色区域中所有像素的和。将计算 C 中特征时，由于黑色与白色区域面积不同，为了保持黑色与白色像素个数相等，因此需要将 Sum 黑乘以 2。由于矩形特征的位置和大小均可以任意改变，所以任何变化都会使得很小的检测窗口包含很多的矩形特征，假设在检测图像大小为 24*24 像素，则生成的矩形特征数量将会有 15 万多，是过完备的。

对于获得不同形式的 Haar-like 特征时，需要计算不同区域的像素之和。这里就需要使用积分图方法进行计算，该方法在求出图像中所有区域像素之和的同时，仅需要对图像进行一次遍历，极大的提高了计算图像特征的效率。积分图最主要操作是在图像中，计算从原点到任一结点所围成的矩形区域中像素之和，并按序存储在一个结构数组中，当需要计算后续其他区域中的像素之和时，仅需要利用公式找到相应元素运算即可。

积分图的构造方式是位置(i,j)处的值$ii(i,j)$表示目标图像中点(i,j)向左

上角到原点所围矩形区域中像素的和，$i(i,j)$表示原始图像，如公式 2-1。

$$ii(i,j)=\sum_{x^{'}\le x,y^{'}\le y} i(x^{'},y^{'}) \tag{2-1}$$

用$s(i,j)$表示某一行像素的累加和，$s(i,-1)$和$ii(-1,\mathrm{j})$初始化为 0，利用公式 2-2 和 2-3 进行计算。

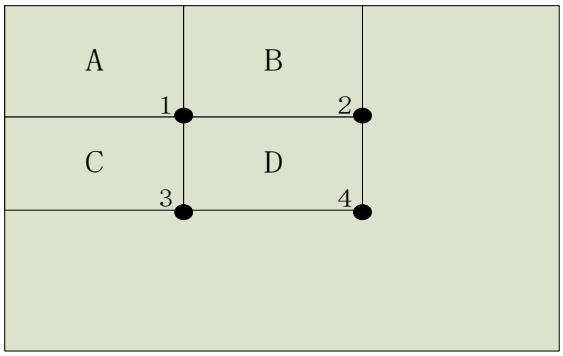

图 2-3　计算积分图

从原点开始对图像进行扫描计算，直到扫描点到达图像右下角的像素，积分图像ii就已经建立好了。

$$s(x,\mathrm{y})=s(x,y-1)+i(x,y) \tag{2-2}$$

$$ii(x,\mathrm{y})=ii(x-1,y)+s(x,y) \tag{2-3}$$

当建立积分图后，计算图中任意矩阵区域的像素之和都可以通过公式进行加减运算得到。如图 2-3 所示，假设矩阵区域四个顶点分别为 1、2、3、4，则区域 D 中的像素和计算方法如公式 2-4。

$$Sum(D)=S(1)+S(4)-S(2)-S(3) \tag{2-4}$$

Lienhart[28]等人在 2002 年进一步扩展了 Haar-like 矩形特征库，增加斜方向特征与中心围绕特征，是该特征具有更好的描述效果，如图 2-4 所示。

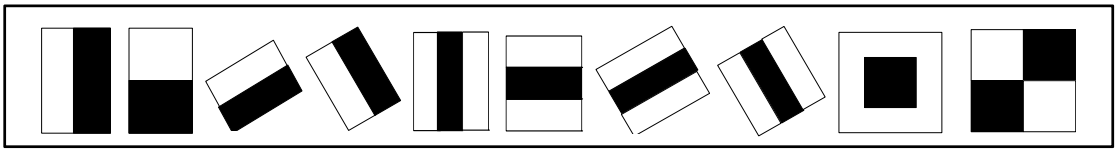

图 2-4　拓展的 Haar-like 特征

2.2.2　HOG 特征

Dalal 等人[1]提出了基于 HOG 特征的人体目标检测算法。HOG 特征是通过在局部区域提取得到其梯度方向的分布来对局部目标进行特征表达，然后对整体目标的形状结构进行表达。由于 HOG 特征是基于对局部区域信息的统计，因此对于目标图像中小的畸变也有很强的抗干扰能力，具有较强的鲁棒性。

HOG 特征的计算主要有五个步骤[29]，具体计算过程如下：

（1） 标准化 gamma 空间和颜色空间

首先为了减少光照的影响，要标准化整幅图像。由于在图像的纹理表现中，局部表层的曝光对图像的特征表达影响较大，因此进行标准化处理能够有效降低图像中阴影和光照的影响。而颜色信息则相对影响较小，仅转换为灰度图进行处理。

Gamma 压缩公式如公式 2-5。

$$I(x,y)=I(x,y)^{gamma} \tag{2-5}$$

（2） 梯度的计算

在计算每个像素的梯度方向值前，需要先计算其横坐标和纵坐标方向的梯度，最后利用公式求得。计算梯度通常进行求导，这样计算不仅可以获得人体轮廓、纹理信息等，光照带来的影响同样也可以进一步削弱。

假设$G_x(x,y)$，$G_y(x,y)$，$H(x,y)$分别表示图像中像素点(x,y)处的水平方向梯度、垂直方向梯度和像素值，则图像中像素点(x,y)的梯度可以通过公式 2-6 与 2-7 计算。

$$G_x(x,y)=H(x+1,y)-H(x-1,y) \tag{2-6}$$

$$G_y(x,y)=H(x,y+1)-H(x,y-1) \tag{2-7}$$

像素点(x,y)处的梯度幅值$G(x,y)$和梯度方向$\alpha(x,y)$如公式 2-8 与 2-9。

$$G(x,y)=\sqrt{G_x(x,y)^2+G_y(x,y)^2} \tag{2-8}$$

$$\alpha(x,y)=\tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{2-9}$$

在计算梯度的运算中，常用[-1,0,1]的矩阵算子与原图进行卷积运算，从而得到横坐标方向的梯度分量，然后用$[-1,0,1]^T$的矩阵算子与原图做相同的卷积操作，得到纵坐标方向的梯度分量。然后再利用上面的公式计算目标像素点的梯度大小与方向。

（3） 构建梯度方向直方图

首先将图像分成小区域（cell），例如每个区域大小为 8*8 像素，在 64*128 的训练图像中，我们就获得 8*16 共 128 个 cell。然后以 4 个田字形 cell 组成一个块（block）。对于每个 cell 中的梯度信息，采用 9 个均匀的通道（bin）直方图进行统计，也就是将 cell 的梯度方向 360 度分为 9 个方向块，就是说设定 20 度是一个通道，0-180 的方向内共分为 9 段，180-360 的方向采用对等角相等的方式进行归类划分。如果一个像素的梯度方向是在 60-80 度范围内，那么在第四个通道的直方图上就做加一操作。这样对 cell 内每个像素用梯度方向在直方图中进行加权投影，就可以得到这个 cell 的梯度方向直方图了，就是这个 cell 对应的 9 维特征向量，而梯度大小就是作为投影的权值的。

（4） 块内归一化梯度直方图

由于图像局部光照的变化，梯度强度的变化范围很大，为了得到更好的检测效果，需要归一化梯度值。归一化可以弱化光照、阴影等带来的影响。如此，一个块的 HOG 特征就是顺序连接内部四个 cell 的特征向量。最后归一化的块描述符就是图像中各区域块的 HOG 特征。

（5） 特征向量的生成

最后，按顺序将所有区域块的 HOG 特征串联收集，形成最终可以用于训练分类器的特征向量。

2.2.3 LBP 特征

在图像分析中常用到纹理的概念，通常表现为亮度或颜色的变化、表面的信

息等，能够将图像的宏观信息和微观结构结合表达。LBP 是一种有效的获得局部纹理信息的简单算法，首先在 1994 年由 Ojala[30,31,32]等提出，该方法通过 LBP 算子来提取灰度图像中局部相邻区域的纹理特征，具有灰度不变性等优点。

LBP 算子类似滤波过程中的模板，通常是 3*3 窗口大小。在计算过程中，先选取目标像素点的灰度值作为比对标准，将该像素点周围八个方向的像素灰度值与选取的像素灰度值做比较，如果中心像素灰度值大，则记为 0，反之则记为 1。以此类推，任意像素周围的 8 个像素点都二值化，然后按照一定顺序将 8 个二值化结果组成一个 8 位二进制数字，这个二进制数字再转为一个无符号整数（0-255），则这个数就是该目标像素点的 LBP 特征值。

由于 8 位二进制数字所表示的整数范围过大，为了提高统计性，提出了一种“等价模式”的处理方式，使得 LBP 算子的模式数量减少。在 8 位二进制结果中，通常只包含 0 与 1 两个数字，利用 0 到 1 与 1 到 0 的跳变次数作为模式分类的标准，例如对于一个 8 位二进制数字 00001111，第四位向第五位转变过程中出现了一次 0 到 1 的跳变，而 11110000 也仅存在一次跳变，因此这两种 LBP 算子属于一类模式。利用这样的分类方式，LBP 算子的模式数量将大大减少，有利于统计，且这种改进不会有任何信息的丢失。通过这种改进，原有的 256 中模式减少到仅有 58 种，在特征向量维数更少的同时，还降低了噪声带来的影响。

(a)　　(b)
原始图像　　LBP图谱

图 2-5　原始图像与 LBP 图谱

通过上述的介绍，可以得知对于图像中的每个像素点都可以根据相应的LBP算子获得一个 LBP 的编码，再通过转化为无符号整数，那么一幅原图进过 LBP 算子的运算后，得到的仍然是一幅图像，图像中的每个像素点值就是原图该像素点的 LBP 值[33]。

在 LBP 的实际应用中，一般不会直接使用 LBP 值作为特征进行分类器训练，而是对原图中划分多个区域，并在区域中进行 LBP 值的直方图统计，最后将直方图的统计结果作为特征进行判别，类似 HOG 特征的提取方式。采用直方图统计的原因在于，LBP 值的获得与位置信息关系密切，对于同样的两幅图像，如果由于图像位置稍有变化，就会发现两幅图像的 LBP 值最终差别很大。因此将图片化成若干的区域，在局部区域内对图像的 LBP 值进行直方图统计，可以降低特征对位置的敏感程度，同时保证了特征的区分性。

例如在一幅 64*128 像素大小的图中，首先对图像使用 3*3 的模板，计算生成 8 为二进制的 LBP 特征值。然后将图像划分区域，每个区域大小为 8*8 像素，则整幅图像分成了 8*16=128 个子区域，每个子区域对 LBP 值进行直方图统计生成 8 位长度的特征值，然后在整幅图像中，串联 128 个子区域的直方图统计值即可描述整幅图像。最后利用相似函数，就可以判别两幅图像之间的区分程度了。

2.2.4 SIFT 特征

SIFT 是一种基于图像局部特征检测的算法，由 David 在 1999 年[34]提出。该方法通过求一幅图像中的关键特征点以及有关尺度和方向的描述得到特征，同时保持图像的尺度和旋转不变性，减少了噪声的干扰，在检测与图像匹配方面获得较好的效果。

SIFT 特征提取[35]分为以下四步：

（1） 构建尺度空间

构建尺度空间是为了模拟图像数据的多尺度特征，是生成 SIFT 特征的初始化运算。在保证尺度变化的前提下，高斯卷积核是目前仅有的线性核。假设 $I(x,y)$

是输入图像，$G(x,y,\sigma)$是尺度可变高斯函数。函数$L(x,y,\sigma)$是一幅图像的尺度空间，计算如公式 2-10。

$$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y) \tag{2-10}$$

其中(x,y)是空间坐标，σ的大小反映图像的平滑程度，大尺度能够反映图像的整体概貌特性，而小尺度则更多表现图像中的细节信息。

高斯函数定义为公式 2-11。

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2} \tag{2-11}$$

为了在尺度空间上能够找到稳定的关键点，利用不同层之间相减的方法生成高斯差分尺度空间（DOG scale-space）。利用不同尺度的高斯差分核与图像卷积生成$D(x,y,\sigma)$，如公式 2-12，其中k为尺度变化系数。

$$\begin{aligned} D(x,y,\sigma)&=\big(G(x,y,k\sigma)-G(x,y,\sigma)\big)*I(x,y)\\ &=L(x,y,k\sigma)-L(x,y,\sigma)\end{aligned} \tag{2-12}$$

总之，构造尺度空间最终目的就是用一幅图像利用公式生成多幅较为模糊的图像，然后将原图缩小，再生成下一层的多幅模糊图像，以此进行操作，则可以构建多尺度的图像金字塔，然后对相邻尺度的图像做差分计算得到 DoG 图像，如图 2-6 所示。

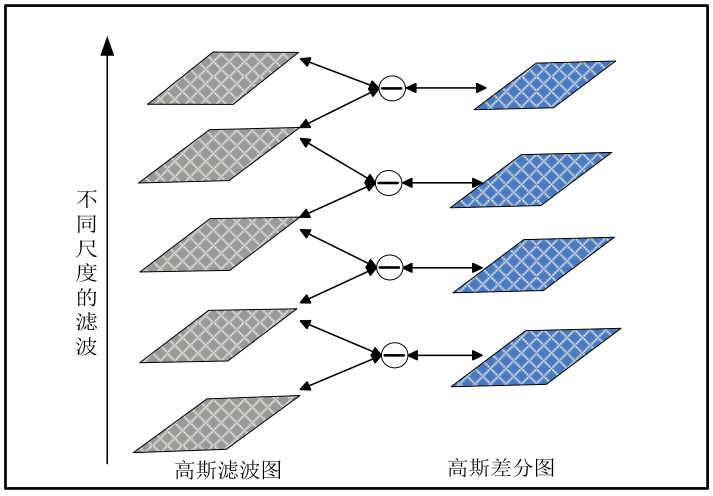

图 2-6　DoG 计算图

（2） 检测局部关键点

在构建的尺度空间中找到极值关键点，要将每一个采样点与附近的点做比较，将采样点附近的 8 个相邻节点以及该采样点上层与下层相对位置的各 9 个结点比较大小，如果采样点比周围的 26 个结点值都大或者小，则该点就是极值点。出现极值点后，则认为该点就是这个尺度下的特征点。

对于上述检测出的候选点，在认定特征点的尺度大小与所在位置之前，要除去对比度较低的其他特征点，还要除去一些不稳定的边缘结点，以增强鲁棒性，降低噪声的影响。

（3） 为关键点赋主方向值

为了保证在特征提取时具有旋转不变特性，因此要转化为关键点的方向来表示。重点在于给每个关键点设置一个具有局部属性的方向来实现。首先在关键特征点周围，计算每个结点在图像中的梯度大小与方向，通过所有点的对比可以获得整体最显著的方向。然后将这个方向作为关键特征点的方向，而剩余的步骤都是按照这个方向进行。这样就可以保证在计算特征时，同一副图像旋转的情况下也可以获得相似的特征。

在关键特征点附近，需要固定一个范围来控制该关键特征点的影响力，当图像尺度越大的时候，控制范围也就越大。在方向控制范围中每个像素点的梯度大小和方向用公式（2-13）和（2-14）计算，因此可以得到另外两幅图，即图像的梯度的幅值图和方向图。

$$m(x,y)=\sqrt{\left(I(x+1,y)-I(x,y)\right)^2+\left(I(x,y+1)-I(x,y)\right)^2}. \tag{2-13}$$

$$\phi(x,y)=\arctan\left(\left(I(x,y+1)-I(x,y)\right)/\left(I(x+1,y)-I(x,y)\right)\right). \tag{2-14}$$

在确认控制范围的主方向后，进行梯度直方图运算还可能存在另外一个或多个峰值相当于主方向峰值的 80%，那么可以适当增加关键特征点，与主关键特征点一起进行判别匹配，这样可以提高匹配的可靠性。

（4） 生成 SIFT 特征。

如图 2-7 所示，以关键特征点为中心获得周围 16×16 的窗口，将其分解为 16

个 4×4 的子窗口。在分别得到的 4×4 子窗口中，利用求导公式计算得到相应的梯度大小与方向，不同于 HOG 特征的是，这里将 360 度的梯度方向分为 8 个段，每隔 45 度为一段来统计窗口中的梯度方向。将关键特征点周围结点的方向信息联合考虑的思想能够提高特征的抗噪能力，同时具备一定的容错性。

这样就可以对每个关键特征点形成一个 4*4*8=128 维的描述子，并且经过上述操作的转换，SIFT 特征向量消除了旋转变化和尺度变化的影响，再将得到的特征做归一化处理，进一步弱化光照的影响。

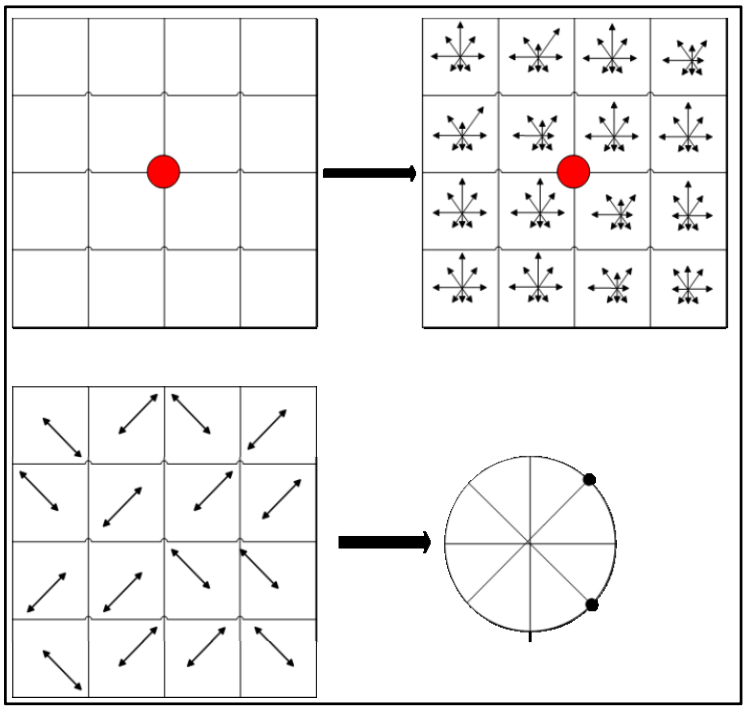

图 2-7　特征点周围的窗口分解，并为每个子窗口创建 8 位直方图

2.3　学习的特征

2.3.1　卷积神经网模型

1984 年基于感受野概念的神经认知机是基于卷积神经网络的一个实例。在卷积神经网中，图像的小部分作为层级结构的最低层的输入，再传输到不同的层，每层通过一个卷积滤波器计算最显著的特征[36]。卷积神经网络中的每一个特征提取层都跟着一个对卷积结果图像做下采样的采样（Pooling）层。采样层能够降

低每一层的特征的维数，同时，使网络在识别时对样本的局部形变具有较高的容忍能力。

CNN 的另一个策略是权值共享。权值共享好处在于网络的输入是图像时会明确的表现出来，图像可以直接输入到整体网络，也显著降低了待求网络权值（参数）的个数。然而因为一个映射面上的神经元的权值是共享的，这样就减少了网络自由参数的数量，在选择网络参数方面的复杂度将大大降低。

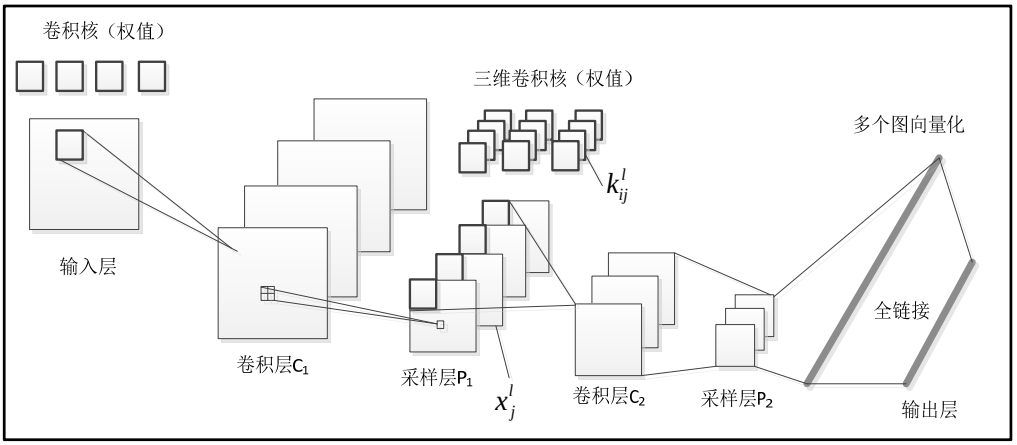

图 2-8　CNN 的结构示意图。C_1 与 C_2 是卷积层，P_1 与 P_2 是采样层

如图 2-8 所示，卷积神经网络是一个分层的结构。*C* 层为卷积层，每个神经元的输入要与上一层的局部感受野连接，通过卷积运算提取局部特征。当局部特征被提取后，它与其他特征的位置关系也将逐渐确定；*P* 层是特征采样层，每个特征映射为一个平面，在平面上所有神经元的权值都相等。在特征映射结构中，由于 sigmoid 函数是影响最小的核函数，因此将采用 sigmoid 函数成为卷积网络的激活函数，是映射的特征具有位移不变性。如图 2-8 所示，输入图像通过和四个卷积核卷积，并将结果通过一个 Sigmoid 函数、加入偏置得到四个卷积层图。卷积层的各个图通过一个采样操作（Pooling）、并将结果通过一个 Sigmoid 函数、加入偏置采样层的四个图。四个采样层图可以看作后面的层的输入图，通过和三个三维卷积核卷积，并将结果通过一个 Sigmoid 函数、加入偏置得到四个卷积层图。卷积层的各个图通过一个采样操作（Pooling）、并将结果通过一个 Sigmoid 函数、加入偏置采样层的三个图。P_2 层各个图的像素值被向量化，并连接成一个

向量输入到一个传统的全链接层，得到输出。

CNN 的训练过程的基本流程是一个误差反向传播（Back-Propagation，BP）算法。为了避免 BP 算法陷入局部最优，人们提出使用 Unsupervised Learning 进行权值初始化。然后，以初始化的权值为初值，进行 BP 算法，计算最终的权值。而当训练数据的规模比较大是，人们发现，直接使用 BP 算法也能获得很好的结果。以下是直接采用 BP 算法学习 CNN 卷积核的学习过程。

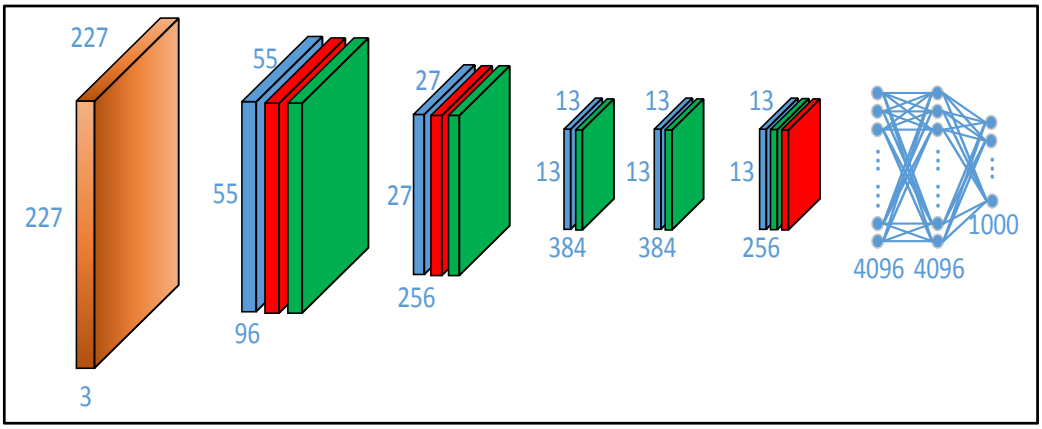

图 2-9　深度卷积神经网络

深度卷积神经网络是一个 8 层的网络，其结构如图 2-9 所示。图中，最左边的棕色的是输入图像，中间五层为前面五层的具体结构，其中，蓝色表示的是卷积层、红色表示的是 pooling 层、绿色表示的是归一化层。在第 5 层 pooling 层之后，特征图都被拉成向量，并连接在一起，形成一个长向量，该向量的每一个元素都构成全链接网络的一个神经元与后面的全链接层相连接，形成一个特征表达。目前这些特征都具有很好的全局表示特性，但局部特征却很弱。

2.3.2　稀疏特征选择

范数是某种距离度量。1 范数是l_p范数中的一种。某一向量的 1 范数是指向量中每一分量的绝对值求和。在实际应用中 2 范数一直都受到学者的青睐，原因在于 1 范数的连续但非光滑性质使得其求解算法比较复杂。直到稀疏表示问题的产生以及压缩感知理论在信号处理领域中的出现，1 范数终于因其优良的性质而得到研究者的青睐。在压缩感知理论中，最早是用 0 范数去分析重构信息的，

然而因为 0 范数的是一个 NP 难的问题，所以该理论使用 1 范数来代替 0 范数，而且理论上已经证明 1 范数可以作为 0 范数的一个近似估计。

此外，在模式识别领域，很多学者应用 1 范数的稀疏特性去判别分类。这里面比较典型的应用有图像重建、人脸识别、图像对齐等。早期的关于 1 范数在模式识别领域应用的文献，要数 John Wright 和 Yi Ma 等人提出的稀疏重构分类（Sparse Representation Classification，SRC）算法。该算法使用 1 范数最小化的稀疏表示识别人脸。该问题是一个多类回归问题，约束方程假设一张测试样本 y 可以由训练样本集 A 进行线性的稀疏表示，目标函数是表示系数 x 的 1 范数最小化。SRC 算法使用 1 范数的目的是希望这种线性表示尽可能的稀疏，最后根据重构误差最小来实现对目标的识别。

$$\begin{aligned} &\min_{x} \quad \|x\|_1 \\ &s.t. \quad y = Ax \end{aligned} \tag{2-15}$$

假设有 k 类不同样本，同一个目标在不同环境、不同光照下的图像构成一个集合 A_i，所有目标构成训练样本集合 $A=\left[A_1, A_2 \cdots A_k\right] \in R^{mxn}$，上述优化模型可等价写成公式 2-16。

$$\begin{aligned} &\tilde{x} = \arg\min_{x} \|x\| \\ &s.t. \qquad y = Ax \end{aligned} \tag{2-16}$$

考虑到噪声与误差的影响，那么公式 2-15 重写成公式 2-17。

$$\begin{aligned} &\tilde{x} = \arg\min_{x} \|x\|_1 \\ &s.t. \ \ \|Ax - y\|_2 \le \varepsilon \end{aligned} \tag{2-17}$$

最后计算残差 $r_i(y) = \left\| y - A\delta_i\left(\widetilde{x_1}\right) \right\|_2$，残差最小的标号则是测试样本 y 类别。

$$class(y) = \arg\min_{i} \ \ \mathrm{r}_i(y) \tag{2-18}$$

从压缩感知到人脸识别问题，这些问题的本质思想，都是通过稀疏表示来实现对物体的识别，属于回归问题。受到上述研究的启发，可以借鉴 SVM、Adaboost 和凹函数支持向量机等分类算法，研究基于了 1 范数最小化的分类问题，结合 1

范数的稀疏性设计检测模型，得到权重向量以及一个人体训练样本的加权特征向量（即权重与人体样本向量的相应维度进行乘积，得到一个同样维度的向量，称为加权特征向量）。当设定一个阈值时，加权特征向量中多数都是小于这个给定的阈值，其余的少数是主要的成份，这一稀疏表示过程可以视为特征选择的过程。实际上，稀疏表示的特征选择的作用是去除冗余信息，尽可能地减少冗余信息将有用的信息淹没的可能性，即可以消除遮挡和噪声对目标的影响，从而提高检测性能[37,38]。

2.4 本章小结

在图像人体检测中，最主要的一个步骤就是特征的提取。特征是图像中提取的能反映图像本质的一些描述子，对于每个提取的特征来说，都有其描述的意义，或是提取目标纹理，或是提取目标形状，都能在一定程度上反映目标的特性。根据近几年研究中经典的特征，本章节从两类特征中选择了一些具有代表性的方法进行重点介绍。

Haar-like特征和LBP特征由于其在局部信息的描述，如颜色和纹理信息等，在人脸检测与识别方面得到广泛应用，但是在人体检测上就没有那么突出的成果。HOG 特征在人体局部梯度信息和轮廓方面也有很好的表征效果，成为目前人体检测方面较好的特征，但由于对其他某些目标的描述与人体目标描述相似，因此在多目标检测中存在一些问题。利用卷积神经网络等方法学习到的特征可以更加有效的使不同类别目标之间的描述最大化，通过使用学习到的特征进行分类器训练，可以进一步提高检测准确率，但是学习的模型通常具有较大的计算量，在计算效率方面性能较差。

第三章　基于稀疏重构与卷积特征方法的人体检测研究

3.1　整体研究框架

近几年，人体检测已经引起了很多研究人员的关注，其广泛的应用场景也迫切需要使得人体检测在性能方面取得更好的效果。在计算机视觉的顶级会议和期刊上，每年都会有很多相关文献，从 2005 年出现的梯度直方图特征，到 2008 年部分可形变的模型框架的应用，再到卷积神经网络在特征提取方面的应用与多通道特征的实现，都在逐步提高对人体外观和形状的描述能力。但是随着当前硬件获取与表达图像的能力增强，单一特征已无法在复杂的环境下达到理想的检测效果。

上一章介绍了在特征提取方面最主要的两类特征：手工设计的特征与基于学习的特征。手工设计的特征在人体检测前期得到很好的发展[39]，其表达能力较强，描述子本身相对简单，速度快，当图像存在畸变时，可以通过相应的操作来弱化畸变或者消除畸变，提高特征表达的稳定性。但是手工设计的特征也存在一些缺陷，如在一种特征描述下，可能会出现几类目标的描述相似度很高，导致不同目标之间的区分性较弱[40]。同时某一种特征无法对所有类目标都有很好的表示与区分能力，例如 Haar-like 只在人脸检测中有较好的效果，在其他目标描述中就不那么理想。而且手工设计的特征提取方法可能会导致丢失部分样本原有的图像信息。由于这些缺陷，基于学习的特征提取方式逐步发展起来。这类方法利用模型自身的学习能力，利用其中某些特定的因子作为特征，不再依靠人为选择的判别标准和规定，方法更具有自主性。由于模型可以根据提供的样本进行学习，因此对于任意目标都会有较好的区分性，同时表达特征的内容更加丰富。但在模型学习过程中，需要经过大量计算，因此运行速度相对手工设计的特征大大降低，而且获得的特征维度高。

在本章的研究中，将考虑两类特征的优点，在实验中结合两类特征进行优势互补，生成结合后的特征描述子进行人体检测，整体实验框架如图 3-1 所示。在

训练阶段，首先进行多通道下的特征提取，如颜色、梯度等，利用多通道下的特征进行三次训练与反例样本挖掘，最后在通道特征的基础上，进行复杂特征的学习，使其表达和描述能力更强，用以训练最终的强分类器。多通道的特征提取属于手工设计的特征，速度快，操作简单。在检测阶段，快速定位图像中的感兴趣区域，获得较少的备选窗口并保证较高的召回率，这样会大大减少最后进行复杂特征提取的窗口数量，在保证利用学习特征提高准确率的基础上，同时保证检测速度较快。在复杂特征学习阶段，本文将采用多种方法实验，主要包括稀疏重构的方法和卷积正交模板的方法。

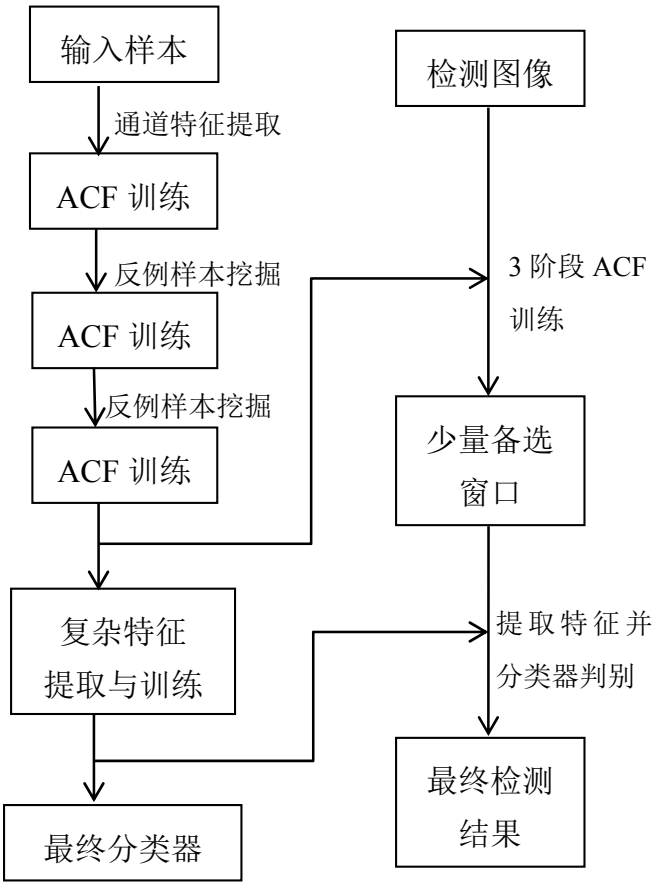

图 3-1　实验整体框架图

3.2　基于集合的多通道特征表达检测方法

基于集合的多通道特征提取方法与整体实验框架如图 3-2 所示[13]。对于给

定的一幅输入图像 I，通过公式$C=\Omega(I)$线性或者非线性的计算规定通道下的像素集合。得到图像 I 在各通道下的图像后，进行图像下采样并做平滑处理，每个通道下的图像大小减半，使得图像的特征维度降低。最后将各采用通道下的特征集合顺序叠加成为 5120 维度的特征，并结合决策树，利用 AdaBoost 分类器进行训练，最后利用对特征上采样与下采样组成特征金字塔，快速的完成对整幅图像的人体检测。

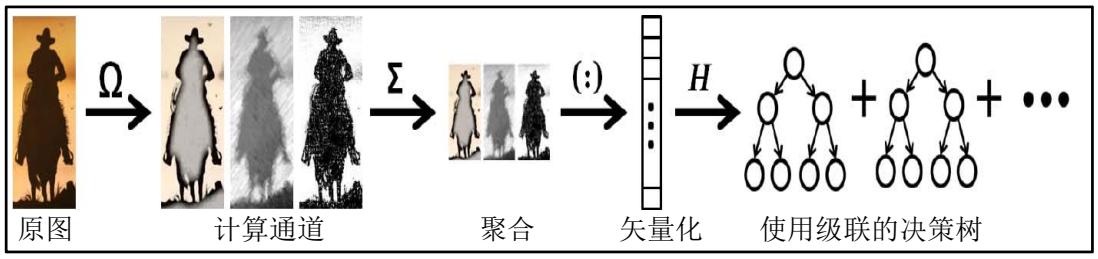

图 3-2　ACF 实验整体框架

3.2.1　通道的选取与表达

最近的研究表明，用积分图计算的方式，对图像多个通道下的像素值进行计算并作为特征，不仅仅对于其他原有手工设计的特征（例如 HOG）特征具有更好的性能，而且可以整合图像中多样的信息资源。而在对参数调整不敏感的同时，采用级联分类器会有更快的检测速度以及准确度。

在实际研究中，对图像进行处理所利用的通道有很多类型，如图 3-3 所示。通道在图像中的意义即为表达显示图像所选择的一种形式，例如彩色通道与灰度通道，这是最简单的两种类型通道。灰度图本身就是图像的一种表达形式，只利用一个像素点就能够表达图像本身。颜色图也是常用的，有 RGB、HSV 和 LUV 等方法。线性滤波也是一种比较直接的方法，还有通过滤波模板得到图像在某一通道下的图像值，例如 Gabor 滤波模板。还有非线性转换方法，例如计算梯度幅度值，Canny 边界等很多方法也可以对图像进行处理表达。除此之外，积分直方图、梯度直方图等也是一类通道类型。利用这些通道对图像处理不仅需要少量代码就可以对图像做标准处理，更重要的是因为很多通道下的计算都非常有效，能够很好从各方面反映图像中的信息。

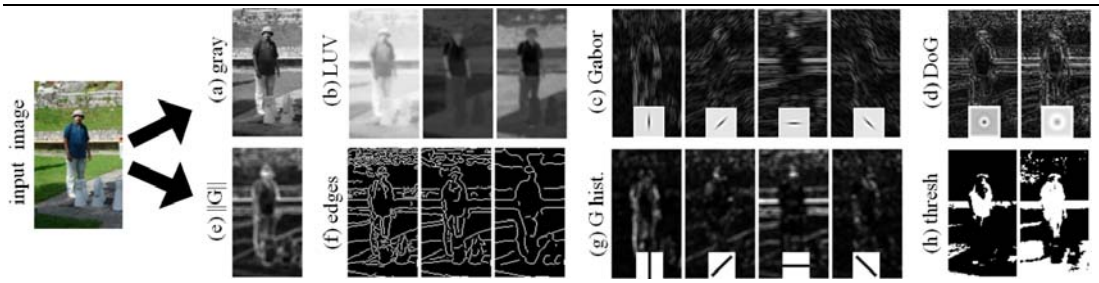

图 3-3　人体图像的多通道信息提取

由 ACF 特征的对比实验中发现，在彩色和灰度通道上，LUV 三通道对图像具有更好的检测效率，同时在其他通道的实验中获得，采用梯度通道以及梯度在六个方向通道上的实验结果性能最好。因此在 ACF 特征提取方法中主要采用对图像 LUV、梯度幅值以及六个梯度方向做处理。

本文将延续 ACF 特征对通道的使用方法，对于一幅图像，每幅图像共提取 10 个通道信息，即颜色 LUV 三通道、梯度幅值通道以及六个梯度方向通道。

由于这十个通道的处理方式是人为规定的，因此属于手工设计特征的一类。通过采用十个通道收集输入图像的信息，不仅可以从更多角度获得图像本身的信息，而且会使得后续利用学习模型得到的特征具有更强区分性，同时提取通道的方法操作简单，运行速率快，为整体算法节省了时间。

3.2.2　分类器的选择

在 ACF 实验框架中，分类器训练部分是结合决策树，使用 AdaBoost 算法训练得到级联分类器。

选择 AdaBoost 分类器是考虑到 ACF 方法中的特征提取方式。ACF 方法在提取特征中，是采用多通道信息的方式进行的，每个通道所代表的信息是不同的，但是单一某个通道的使用并不会带来很好的效果，有效的综合各通道的信息同时又不将所有信息混为一体来评测结果是一个难点。例如在一场比赛中，参与队伍的整体实力、近期表现、球员个人能力以及比赛场地等都会影响比赛结果，但是其中任一因素都不足以决定比赛，如何能够结合所有因素同时做出很好的预测结果呢？

AdaBoost 算法提供了一个很好的方法，它的核心内容是对于同一个训练集，从那些有一定价值的角度训练得到一个具有单一评价标准的性能较弱的分类器。以此类推，不同的角度获得不同的弱分类器，然后把这些弱分类器串联结合，建立一个判别性能更强的最终分类器。实际中，只要在对应的角度将每个弱分类器训练好，最终的强分类器就可以达到任意精度。除此之外，当存在新的有价值的信息时，可以再训练新的弱分类器加入到最终分类器中，因此该方法还具有很好的泛化能力。

AdaBoost 算法的实现需要依靠改变数据的分布状态，算法的关键是不同样本权值的更新，根据每次训练集中所有样本的判别正确与否，以及前一次在总体角度上的判别分类准确率，来确定如何修改每个样本的新的权值。当样本被分错的时候，其权值会增大，反之则会减小，从而逐步将训练的关键集中到难样本上。将修改过权值的样本送到下次分类器训练中，最后把训练得到的弱分类器串联结合起来，组建成为了最强的决策分类器。

对于 AdaBoost 分类器中，每一个弱分类器是利用决策树进行的。决策树通过自顶向下构造的方式从根节点排列到某个叶子节点来分类实例，在实例中不同的分类则在叶子节点上表现出来。决策树中的每一个节点将对实例进行某一方面的测试，并且根据不同的测试结果，相继分布在后续的子节点上，测试所产生的所有可能都会出现在节点中。分类实例的方法是从整棵树的根节点开始，以此对当前遇到的节点所代表的属性进行测试，依据测试的结果，按照后续分配的节点下移，并进行下一个属性的测试。

在 ACF 实验框架中，以一颗决策树的学习作为一个弱分类器，利用提取的特征，先后进行四个阶段的训练，每次训练使用的决策树数量翻倍增加，分别为 32、128、512、2048 棵决策树，最终训练得到一个具有 2048 个弱分类器组成的强级联分类器，如图 3-4 所示。在每次训练结束后，进行反例样本挖掘，逐步得到更为精确的检测窗口。

本文的研究是建立在 ACF 实验框架的基础上进行改进，在特征提取方面，

同样是选择了多通道信息的提取，因此为了更好的保留不同通道的信息，本文在分类器方面不做更多研究，依旧采用 AdaBoost 算法训练分类器。

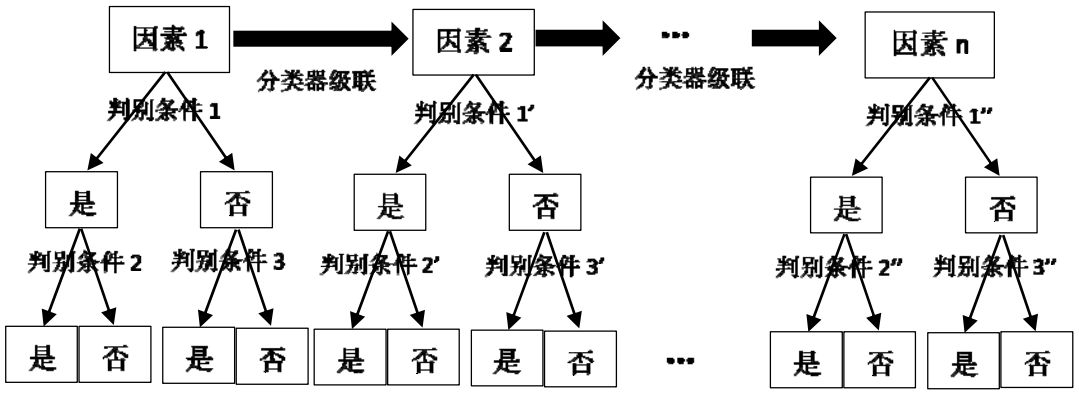

图 3-4　决策树构成的 AdaBoost 分类器

通过介绍多通道特征的提取，以下将从两个部分介绍结合两类特征的实验方法。两种方法都是以多通道特征作为手工设计的特征，其中一种方法是结合稀疏编码学习模型生成稀疏重构特征，另一种方法是结合主成分分析法等学习生成正交滤波模板，然后卷积正交模板生成特征。

3.3　基于多通道的稀疏重构特征表达方法

3.3.1　稀疏重构的特征表达

稀疏编码是一种人工神经网络方法，这种方法具有空间局部性和方向性，通过多位数据处理方法进行编码，能够最后获得少量的分量，同时更好的呈现原信号，是一种自适应的图像统计方法。

随着学习特征的研究逐步开展，稀疏编码方法也慢慢应用在特征表达方面[41,42]。从压缩感知到人脸识别问题，本质上都是利用稀疏表达来描述物体，通过反复迭代计算得到权重向量系数以及通过物体本身训练得到的字典向量集，即权重向量系数与字典向量集进行乘机可以得到与表达物体同等维度的向量，而在其中，权重向量系数就是一个稀疏表达，这个过程可以视为特征提取过程。在实际中，虽然最后获得的稀疏表达很多值为零，但是从本质上它去除了冗余信息，

尽可能的让有效信息在冗余信息中凸显出来，减少噪声以及遮挡对目标的影响。

稀疏编码的方法蕴含了结构相似度的原理，对于原图像中的结构信息能够很好的保留下来，并通过稀疏重构的方式获得稀疏编码特征。在稀疏编码的方法中，输入图像可以通过基函数线性的表达出来，并通过计算最小的均方差得到线性表达的系数，从而使得重构出来的图像尽可能与原图像相似。为方便叙述，我们将原图像按列排成一个 N 维列向量，用 I 表示，$I_i\left(i=1,...,N\right)$ 表示每一个像素点。其中，用于编码和重构的基函数矩阵用 D 表示，大小为 $N*M$ ，其中每一个 N 维列向量用 $d_i\left(i=1,...,M\right)$ 表示，用 M 维列向量 X 表示原图像对应的“响应”，$x_i\left(i=1,...,M\right)$ 表示每一个响应值。进行重构后的图像用 N 维列向量 Y 来表示。

$$Y=DX=\sum_{k=1}^{M}x_k d_k \tag{3-1}$$

$Y_i\left(i=1,...,N\right)$ 表示重构图像的每一个像素点。我们引入稀疏重构误差最小化目标函数中，目标函数如公式 3-2。

$$E^*=\arg\min_x \left\|I-DX\right\|_F^2+\lambda \mathrm{s}\left(x\right) \tag{3-2}$$

其中 $\mathrm{s}\left(x\right)$ 是零范式，要求 $s.t.\forall i,\left\|x_i\right\|_0\le K$ ， K 是预定义的稀疏程度。当计算得到的误差值越小，则重构后的图像与原图像越相似，保留原图像中的信息更多，而稀疏程度更高时，目标函数就会更小。根据有效编码原理，应使得重构误差和尽量小，使响应个数也尽量少。

在得到最优化目标函数 E^* 后，即得到最有的响应值 X ，则原图像的稀疏编码特征就是 X 。这里描述的是一种相对简单直接的特征表达方法，即直接利用稀疏编码作为特征使用。除此之外，还可以将稀疏编码与基函数矩阵相乘，利用重构后的数据作为新的特征表达，同时保留稀疏性的优点，去除冗余信息。

3.3.2　实验方法

在前面分析手工设计的特征与基于学习的特征各自的优势与缺陷后，将继

续研究两类特征融合后所产生的新描述子，对于人体检测带来的效果。

在手工设计的特征中，根据对不同通道下实验可知，选择在图像六个梯度方向、梯度模值以及 LUV 等十个通道下提取相应的特征，各通道下的提取结果如图 3-5 所示。

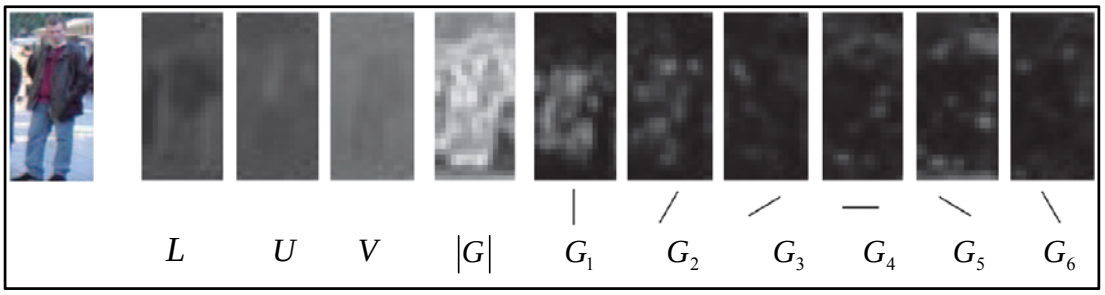

图 3-5　手工设计特征方法下的多通道特征提取

在学习的特征中，将利用稀疏编码模型，对多通道下提取的特征进行运算编码，然后利用编码进行稀疏重构，提取重构后的图像作为特征，最后训练分类器进行实验。图 3-6 为实验流程图。

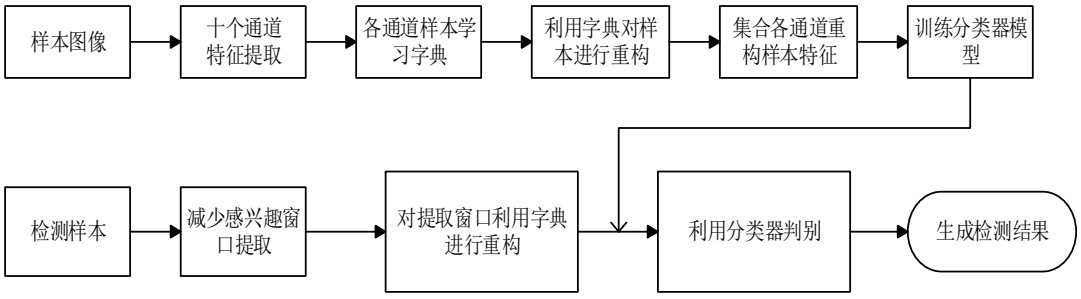

图 3-6　实验流程图

在训练阶段，首先在正例样本中提取规定的十个通道下的特征表达，然后利用 AdaBoost 分类器进行三次训练并挖掘难反例。在最后一个阶段，进行复杂特征提取。首先在每一个通道下的特征图像中，使用稀疏模型进行正例样本中字典的学习与正例样本稀疏表达的获得。由于正例样本集是由多种不同姿态、不同环境以及带有部分遮挡的人体图像，因此使用稀疏模型学习到的字典中元素是描述人体不同角度的元素，适当的结合字典中部分元素，即可表示出不同类型的人体。因此，在经过稀疏运算后，每一幅图像都可以根据字典获得一种对应的稀疏编码表达。再将稀疏表达与字典进行乘积可以重构生成一幅新的正例图像，将图

像的像素链接起来则可以作为该通道下最后的特征表达。当所有通道下的图像都进行稀疏重构表达后，串联不同通道的特征表达即可获得最后的特征描述子。

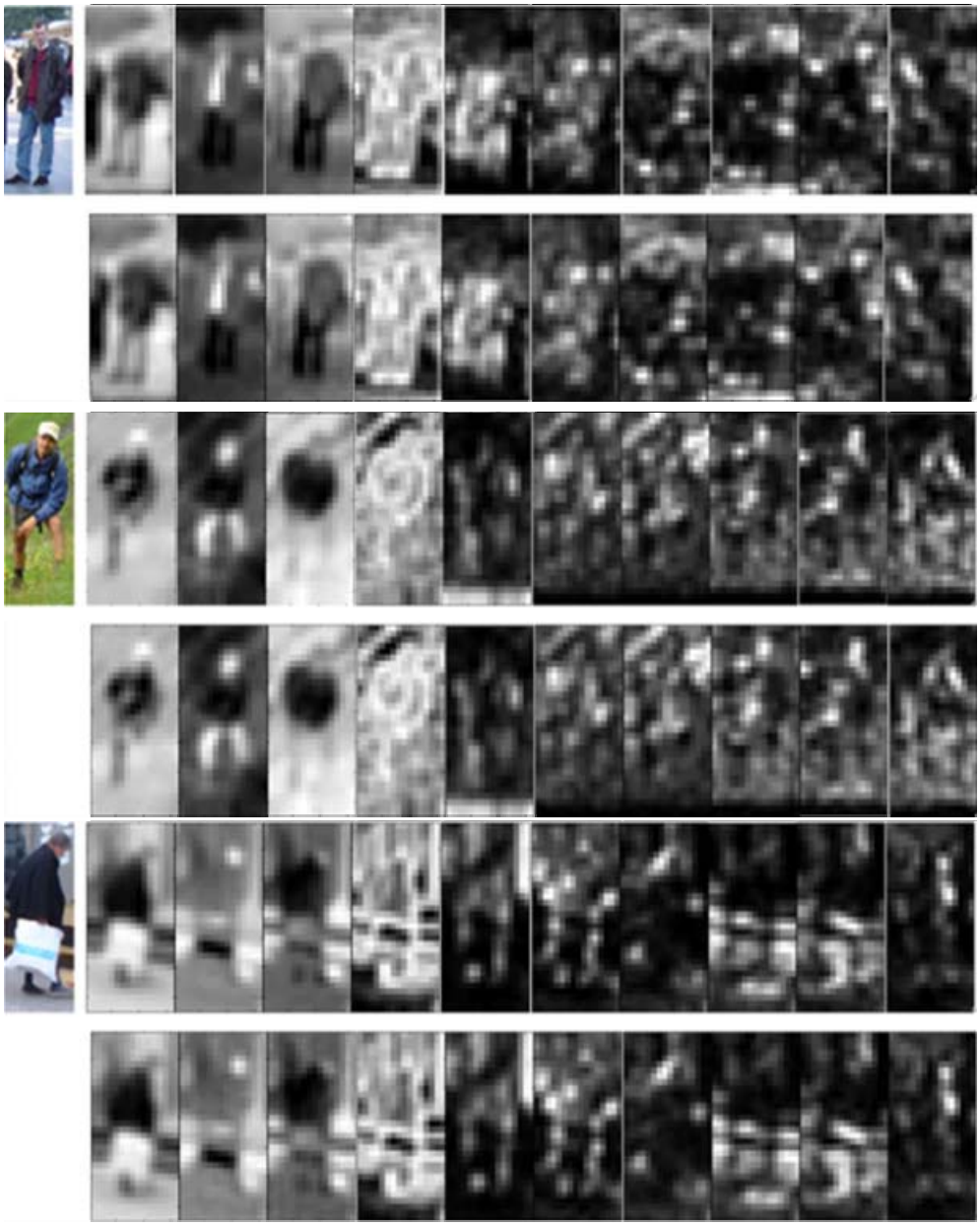

图 3-7　不同通道下原样本（第一层）和重构后样本对比图（第二层）

那么利用稀疏编码表达进行重构有什么好处呢？重构最主要的原因是，利用字典中元素组合，在尽可能恢复原图的情况下，将弱化源图像中光照以及遮挡的影响。由于人体的光照变化以及遮挡并不是存在于每一幅图像中，而且遮挡的

位置也会不同，同时这些因素并不会成为图像中的主要成分，而在利用稀疏模型进行字典学习过程中，这些遮挡与光照的影响已被编码运算分解开，并利用模型的稀疏性将其过滤掉。因此在重构后的图像中，那些曾经带有遮挡以及强光照变化的图像变得模糊，在保留人体整体形态的同时，弱化了遮挡与光照。图 3-7 为实验中重构前后对比图，图中最左边图像为原样本图像，第一层图像为重构前各通道下的特征表达，第二层为重构后的各通道下特征表达。

3.3.3 实验结果与分析

在本文的实验中，将以 ACF 实验结果作为基础进行两种实验测试。一种是直接利用图像获得的稀疏编码表达作为最后的特征进行复杂分类器训练，另一种则是利用重构后的图像进行特征表达。实验的数据集为 INRIA 数据集和 Caltech 数据集。

（1） 直接利用稀疏编码作为特征进行实验

在第一阶段得到原始图像的多通道下图像后，本文的研究首先利用对多通道下图像做稀疏编码，在每个通道下对样本图像做字典学习，最后由学习到的字典编码直接作为最后特征表达输入到最后的分类器训练中。

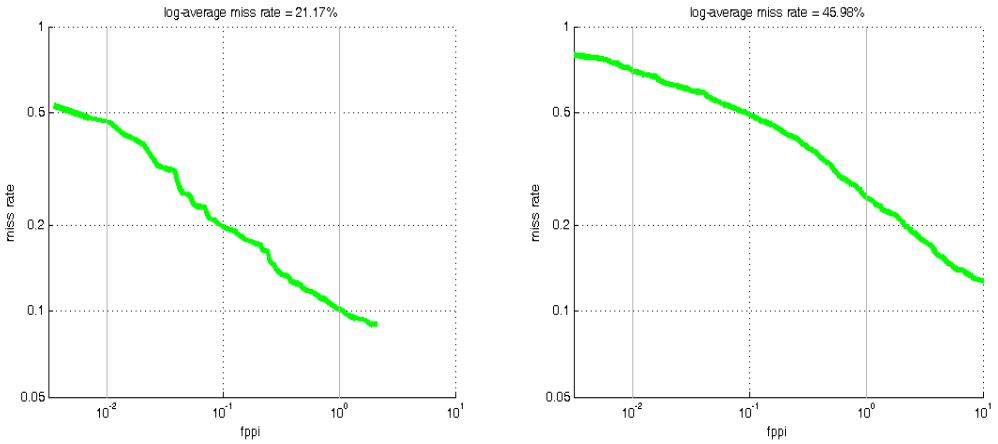

图 3-8 编码特征在 INRIA 与 Caltech 上的性能曲线

如图 3-8 所示为直接使用字典编码作为特征在两个数据集上的性能曲线。在原 ACF 实验框架下的特征表达在 INRIA 数据集上的 missrate 为 17.28%，在

Caltech 数据集上的 missrate 为 43.94%，而经过结合稀疏模型，利用图像稀疏表达作为特征，在两个数据集上的 missrate 分别为 21.17%与 45.98%，结果表明直接使用字典编码作为特征输入到最后的分类器训练中并不能提升性能。

直接利用稀疏编码作为特征进行分类器学习的方法，与 ACF 实验方法在数据集上的检测对比效果如图 3-9 所示。

图 3-9　利用编码作为特征的实验结果（左）与 ACF 实验结果（右）

这种方法的检测性能下降，原因应该主要分两方面：一是直接利用字典编码作为图像最后特征是属于一种新类别的特征。由于稀疏重构的重点是重构输入图像，原始图像在经过多通道下处理后，虽然表现形式不同，但是图像表达内容会保留下来，在之后重构的过程中，各通道下的稀疏编码已失去了不同通道的区分性，这样就类似于在单一通道下做字典学习并编码，编码特征中包含信息量较少，因此无法有效的提升检测性能。第二方面是这类条件下的稀疏编码没有较强的区分性。

（2） 利用稀疏重构后图像做特征进行实验

为了更好的保证原通道信息的保留，本研究进一步对稀疏编码做计算，利用稀疏编码和学习的字典相乘，得到新重构而成的各通道下的图像表达，再使用新重构图像作为特征输入到最后一级分类器训练中，得到性能图如下：

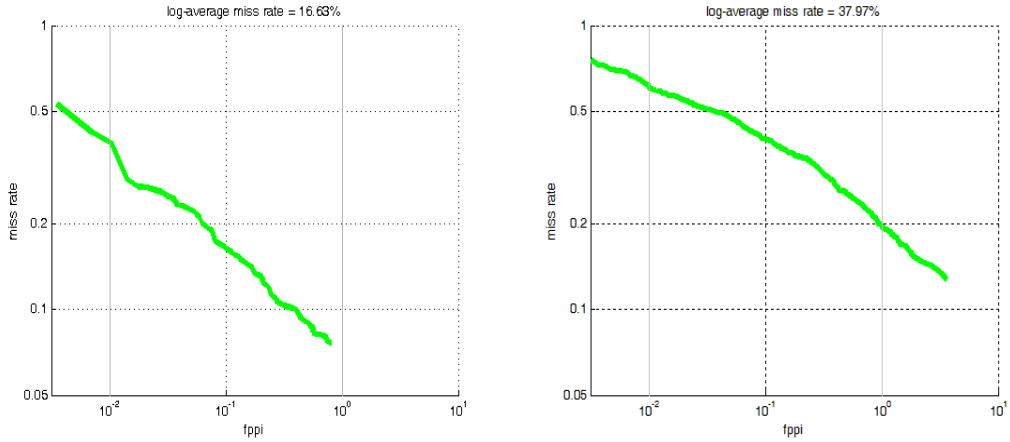

图 3-10　重构特征在 INRIA 与 Caltech 上的性能曲线

实验结果发现，利用稀疏编码对原各通道下样本图像进行重构的特征提取方法在检测结果上较 ACF 实验框架有所提升。该方法在数据集上的检测效果如图 3-11 所示。

图 3-11　实验结果图

进行重构得到的特征在检测性能上有所提升，分析的原因可能是由于字典

学习所使用的样本是正反例图像，而正例与反例的区别主要在于某些元素上不同，在字典学习的过程中，组成正反例的各个元素会表现为字典中不同的不同模板，当利用学习到的稀疏编码与字典正交重构图像时，该重构图像是由字典中不同模板组合而成，因此图像整体会变模糊，但是图像轮廓及外形则不会发生较大变化。这样由重构带来了一点好处就是，虽然重构图像变模糊，但是也同时弱化了光照，尤其是弱化了遮挡带来的影响。所以，在最后检测性能上会比 ACF 中只用通道特征要更好更有效。

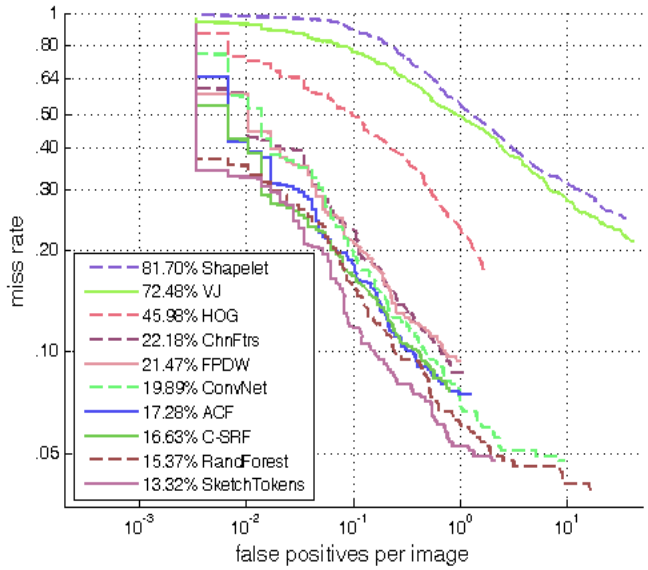

图 3-12　常用特征在 INRIA 上的性能对比

图 3-12 为常用特征在 INRIA 数据集上的对比曲线，其中本节提出的基于多通道的稀疏重构特征（Channel - Sparse Reconstruction Feature，C-SRF）有较好的效果。由此可知，将手工设计的特征与基于学习的特征进行结合的方法，在一定程度上可以有效的提高检测效果。

3.4　基于多通道的卷积正交模板特征表达方法

3.4.1　主成分分析方法

主成分分析法是从多个变量中选出少数具有代表性变量的统计分析方法。

例如某个变量在样本中彼此差距很小，用这个变量作为特征，贡献率低，因此将变量映射到另一个维度，使新得到的变量方差变大，那么转换后的变量就是关键变量，而计算量也相对减小。统计分析中，不同的变量代表不同的信息内容，不同变量之间可能有一定相关性，通过进行维度转换，使生成的新变量之间都是正交的，去除了变量间相关性，减少冗余信息的统计。

主成分分析法的计算通常具有以下五步：

（1） 对原始数据做标准化处理。

提取 n 个样本 $x_i=(x_{i1},x_{i2},...,x_{ip})^T$，$i=1,2,...,n$，每个样本是一个 p 维的随机变量 $x=(\mathrm{x}_1,x_2,...,x_p)^T$，$n>p$，对样本进行标准化变换。

$$Z_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j},\quad i=1,2,...,n,\quad j=1,2,...,p \tag{3-3}$$

其中 $\bar{x}_j=\frac{\sum_{i=1}^{n}x_{ij}}{n}$，$s_j^2=\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}$，计算得到标准化数据 Z。

（2） 对 Z 求相关系数矩阵

$r_{ij}(i,j=1,2,...,p)$ 为原变量 Z_i 与 Z_j 的相关系数，计算如公式 3-4。

$$r_{ij}=\frac{\sum_{k=1}^{n}z_{kj}\ z_{kj}}{n-1} \tag{3-4}$$

相关矩阵为 $R=\left[r_{ij}\right]_p\ xp=\frac{Z^TZ}{n-1}$

（3） 计算相关矩阵的特征值与特征向量

解特征方程 $\left|\lambda I-R\right|=0$，求出 p 个特征值，并按其大小进行排列。

$$\lambda_1\geq\lambda_2\geq\Lambda\geq\lambda_p\geq0 \tag{3-5}$$

同时解方程组 $Rb=\lambda_j b$，并要求 $\left\|b_j\right\|=1$，即得到单位特征向量 b_j^o。

（4） 计算各成分的贡献值，并将标准化后的变量转换为主成分

各成分的贡献率如公式 3-6 所示。

$$g=\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k},\quad i=1,2,...,p \tag{3-6}$$

累计的贡献率如公式 3-7 所示。

$$h=\frac{\sum_{k=1}^{m}\lambda_m}{\sum_{k=1}^{p}\lambda_k},\quad m=1,2,...,p \tag{3-7}$$

一般取累计贡献率为 85%-95%的特征值，当$h\geq 85\%$时，取相应的特征向量b_j^o，并转换为主成分U。

$$U_{ij}=z_i^T b_j^o,\quad j=1,2,...,m \tag{3-8}$$

其中U_1成为第一主成分，U_2是第二主成分，U_p是第p个主成分。

（5） 对各主成分进行分析统计，评价每个主成分的贡献率。

主成分分析方法最终将彼此相关的指标变量转化为彼此不相关的指标变量，同时减少了变量的个数，将原来意义单一的变量转化为带有综合意义的指标。

3.4.2 实验方法

在生成复杂特征描述子的实验中，本文结合主成分分析法[43]等，通过这些方法模型学习获得正交模板并进行卷积运算，生成最后的复杂描述子。

在训练阶段中，同样按照 ACF 实验框架，对每幅图像取十个通道下的图像表达，进行三次训练后生成一个较为弱化的一个分类器，然后用通过正交方法生成的模板与样本图像做卷积运算，得到最后的复杂特征，训练最后一级分类器。在检测阶段中，检测图像经过弱化的分类器可以生成较少的备选窗口，窗口数量比扫窗方法得到的窗口数量减少很多。这些备选窗口对于目标人体有着很高的召回率，但同时也包含了很多非人的窗口。然后，利用正交模板与备选窗口做卷积运算，生成最后的复杂特征，进入最后一级分类器进行判别。

本文在利用正交方法生成卷积模板中，主要使用主成分分析法。对于给定的N副样本图像$I=\{I_i\},i=1,...,N$，图像大小为$w\times h$，可以计算得到多通道下的表达$M=\{M_{i,k}\}$，$i=1,...,N$，$k=1,...,10$，k代表通道数量。对一幅图像中的每个像

素取5×5大小的块，并按顺序将每个块的像素串联起来形成一个长度为 25 的列向量。因此对于一副图像我们将获得 $w\times h$ 个向量，表示如公式 3-9。

$$X_{i,k}=\left[x_{1,k},x_{2,k},...x_{wh,k}\right]\in R^{k_1k_2\times wh} \tag{3-9}$$

之后利用 PCA 方法进行运算，PCA 使用一种正交变换方法将一组相关变量转化为一组具有线性无关的值，同时对于转化后的正交向量重构误差最小化。因此在经过 PCA 计算后，每幅图像上提取的列向量个数将减少，实现降维。按照变换后各值的重要性，本文在实验中取出前八个值代表整体，所以在每个通道下，将获得 8 个5×5大小的滤波模板块 $F=\{F_{l,k}\},l=1,...8,k=1,...10$ 。以此类推，在所有通道下获得的模板示图如 3-13[44]：

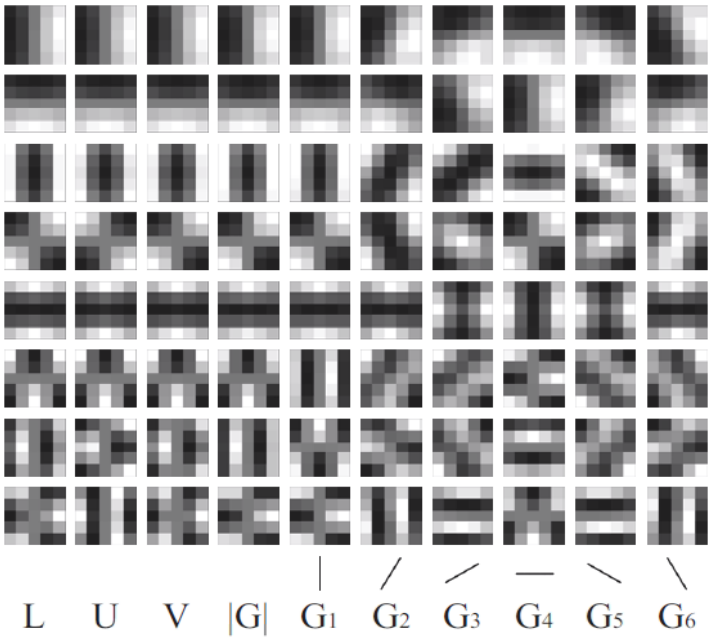

图 3-13　十通道下的 PCA 滤波模板，每列一组包含 8 个特征向量

在学习得到正交滤波模板后，下一步进行卷积运算。首先在某一通道下，将 M_k 与上述得到的正交滤波模板做卷积，即

$$Layer_{c,l}=F_{k,l}\circ M_k \tag{3-10}$$

其中 $F_{k,l}$ 是在第 k 个通道下的第 l 个 PCA 滤波模板，而 M_k 是第 k 个通道下的特征图。

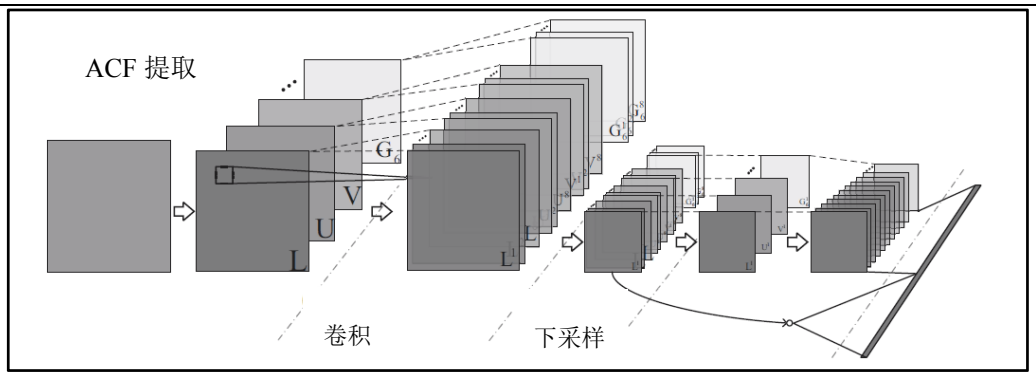

图 3-14　各通道下卷积滤波模板的示意图

在卷积神经网中，下采样是一种常用的操作手段，旨在降低特征维度，增强鲁棒性。在卷积操作中，下采样就是每隔两个像素取一个值，将原图的长与宽各减少一倍。在上述进行卷积操作后，就对结果做下采样处理，并在同一通道下，将卷积生成的多幅图像合并为一，作为最后的特征表达。如图 3-14[44]所示为各通道下卷积滤波模板的流程。

3.4.3　实验结果与分析

利用主成分分析法学习得到正交模板后，将图像与模板进行卷积操作得到复杂特征，并最后训练生成分类器。在测试过程中，将在常用的人体检测数据集 INRIA 与 Caltech 数据集上进行实验。如图 3-15 所示，在 INRIA 上的 missrate 从 17.28%降低到 14.24%，在 Caltech 上的 missrate 从 45.98%降低到 32.84%，检测性能都得到了提升。

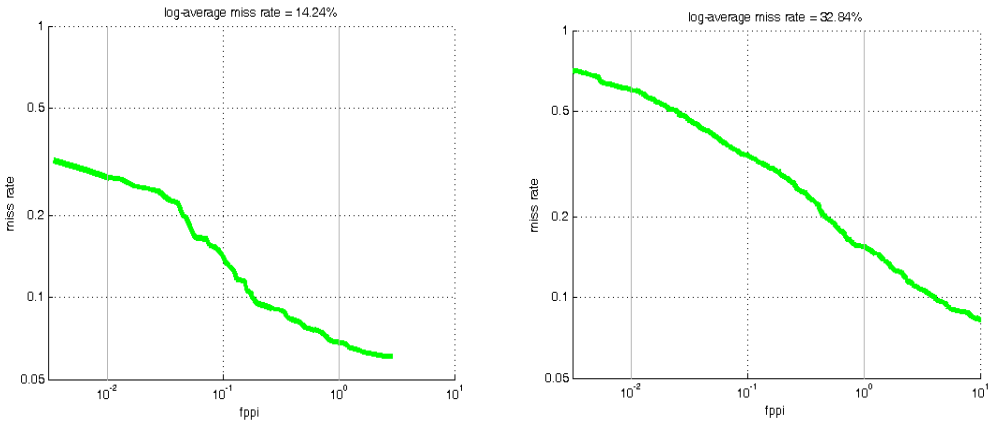

图 3-15　卷积正交模板特征在 INRIA 与 Caltech 上的性能曲线

实验结果显示，对正交模板进行卷积生成复杂特征的方法，最终将有助于检测性能的提升。图 3-16 为检测结果示意图，可以看到对于每一幅检测样本图中，分类器将对图中可能检测到的感兴趣区域做方框标定出来，这里的感兴趣区域即分类器认为该区域为人体。在检测的同时，分类器对检测区域进行打分，当得分为正，表明分类器认为该区域为人体，得分为负表明该区域为非人体，得分越高说明该区域为人体的概率越高。在结果图中可以看到，在与 ACF 方法结果图 3-9 相比，通过卷积正交模板的结果有更好的性能，提高了人体区域的得分，同时减少了 ACF 方法中检测错误的区域。

图 3-16　检测结果示意图

在对正交模板进行卷积的方法中，生成模板的大小与模板的个数也是影响实验结果的两个主要参数，因此本文做实验进行参数对比。卷积正交模板是对卷积图像中的细节进行分析的操作，小的模板可以表示图像中的横、竖或者其他简单的细节信息，但对区域信息没有很好的表达，而大的模板则可以表示一些局部区域的信息，但对细节信息描述较弱。因此我们在保持实验方法流程不变，对模板大小分别设置为 3*3、5*5、7*7，模板个数设置为 5-8 个，以此进行实验，表 3-1 为在 INRIA 数据集上的 missrate 结果对比。

表 3-1　参数对比结果

	5	**6**	**7**	**8**
3*3	16.33	16.16	14.5	16.27
5*5	17.27	16.21	14.62	14.24
7*7	17.58	18.1	17.71	18.16

通过对比发现，在模板大小为 5*5 像素以及模板个数为 8 的时候，性能较好。说明在卷积中，我们需要将细节信息与局部区域信息都考虑进去，将对检测有好的帮助。而且适当增加正交模板的个数，也对检测结果有提升。

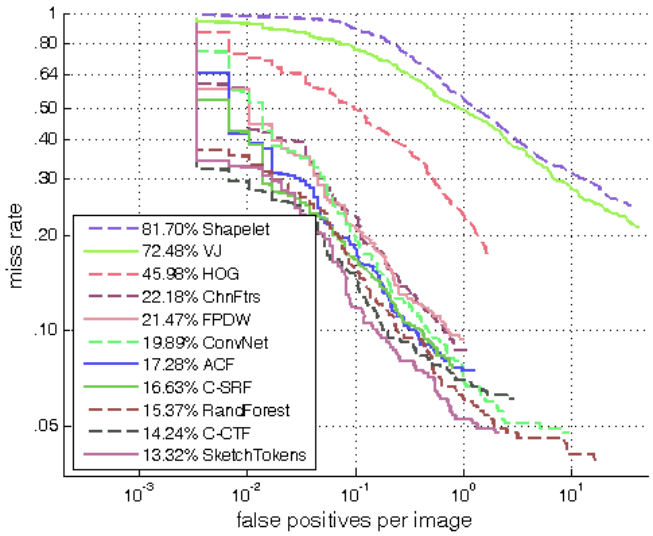

图 3-17　常用特征在 INRIA 上的性能曲线

如图 3-17 所示，本节提出的基于多通道的卷积正交模板特征（C-CTF）与已有的常用特征相对，在性能上有了一定的提升。利用主成分分析等方法生成正

交的滤波模板，能够有效的去除冗余信息，使得在后续卷积运算中，图像特征的区分性更强，同时也表明有效的结合两类特征能够提高检测效率。

3.5　本章小结

本章在多通道特征的基础上，利用稀疏模型以及卷积正交模板的方法得到更加复杂的特征描述子，在实验结果上都有了一定的提升，说明结合手工设计的特征与基于学习的特征，将有助于增强特征的表达能力，同时利用实验，得到卷积正交模板也会有助于提高特征的区分能力，提高检测结果性能。

第四章　高分辨率数据在人体检测中的研究

4.1　问题分析

人体检测在计算机视觉领域已经进行了很多年研究，研究的方法主要分为两步，一是利用输入图像提取特征，二是将获得的特征用于训练分类器。因此，多年来，在人体检测方面的研究重点都集中在如何获得更好的特征提取方法与训练更加有效的分类器，同时通常用于评测检测性能好坏的数据集保持仅有的几个种类。

随着社会的进步，人体检测的应用范围越来越广泛，当前使用的硬件设备已经能够拍摄较高分辨率的图像与视频，然而用于评测人体检测方法性能的数据集仍然是相对低分辨率的，与现实拍摄数据的清晰度与图像视频的大小有一定距离。

近几年人体检测的性能在逐步提高，但是也慢慢遇到瓶颈，当检测性能提高到某一范围内，再次获得提升的难度越来越大。在特征提取方面，从单一手工设计的特征方法到卷积神经网络等多种学习模型的跨领域应用；在分类器训练方面，从原来一种分类器到多类分类器联合，然而在检测性能方面却没有特别大的提升，检测效率还是无法满足当前实际应用的最低需求。

因此，本文使用现有的特征提取方法与分类器方法，采集高分辨率数据集并在此数据集上进行评测实验，探索高分辨率数据是否会对提升检测性能有好的帮助。

人体检测中的特征提取以及训练分类器，其最主要的目的都是希望尽可能的使人体形态在图像中更加明显，进而定位。由于高清数据集相比较现有常用的人体检测数据集，分辨率与图像大小都有很大的提升，视频图像中的人体会相对显示得更加清晰，因此，提高图像本身质量有可能在一定程度上缓解对特征提取的要求。本节内容就将对现有常用人体检测数据集与新拍摄的高分辨率数据集进行对比分析，并进行实验。

4.2 人体检测数据库

表 4-1 列出了现有的一些主要人体检测数据集。通过表格看出根据获取数据的方式不同，当前数据集主要分为静态图片与动态视频两种。INRIA 数据库自 2005 年发布以来，就成为了评测人体检测技术的主流数据集，极大的推动了人体检测技术的发展。数据集 PASCAL VOC 2007 是多目标检测数据集的典范，然而很多好的特征描述子与分类器算法在该数据集上都没有取得令人满意的效果，说明在人体检测方面还有很大的研究空间。INRIA 数据集是专门用于进行人体检测的，而 Caltech 数据集则是利用车辆上安装的摄像头拍摄而成，说明人体检测应用在移动端的急切需求。

表 4-1 人体检测公开数据集表

数据集	训练集	测试集	注解
INRIA People Dataset	2416(64*128) / 1218(含标定)	1132(64*128) / 453(含标定)	1) 彩色图片 2) 背景复杂
Caltach Pedestrian Dataset	250,000 帧, 350,000 bounding boxes 2300 行人（标定）		1) 驾驶员视角 2) 视频帧序列
TUD-Brussels Pedestrian Dataset	508 幅 640*480 的图片，共 1326 个行人		多尺度、多视角的行人
VOC2007 dataset	正样本 4096 /负样本 3003	4528 行人	1) 彩色图像 2) 行人检测最具挑战和影响
TUD Multi-view Pedestrians Dataset	4732 幅彩色图片（含标定数据）	250 幅彩色图片	1) 图片尺寸不一 2) 含部位标定 3) 共 8 个视角
SDL Pedestrian Dataset	正面 1000 幅 / 侧面 3050 幅 / 多视角 7550 幅	140 幅彩色图片 /258 幅多视角图片	包含正面/侧面/多视角 64*128 像素（训练）

在本文的研究中，现有常用的人体检测数据集为 INRIA 数据集和 Caltech 数据集，以下对两个数据集与本文新拍摄标定的高分辨率数据集进行分析。

（1） INRIA 数据集

该数据集是由 Dalal 在 2005 年 CVPR 提出 HOG 特征时一并公开的，主要由两个部分组成：原始图像以及相应的标定注释文件。

数据集包含了几个不同来源的图像：从格拉茨图片集中提取、从一些较高分辨率中图像剪裁以及从谷歌网络图像中获得。在标定方面，该数据集只标定直立人，其他姿态的人体不选取。

INRIA 数据集的训练集正例样本数量为 614 幅，反例样本为 1218 幅，测试集仅包含 288 幅图像，并且每一幅图像分辨率不同，大约 800*600 左右。

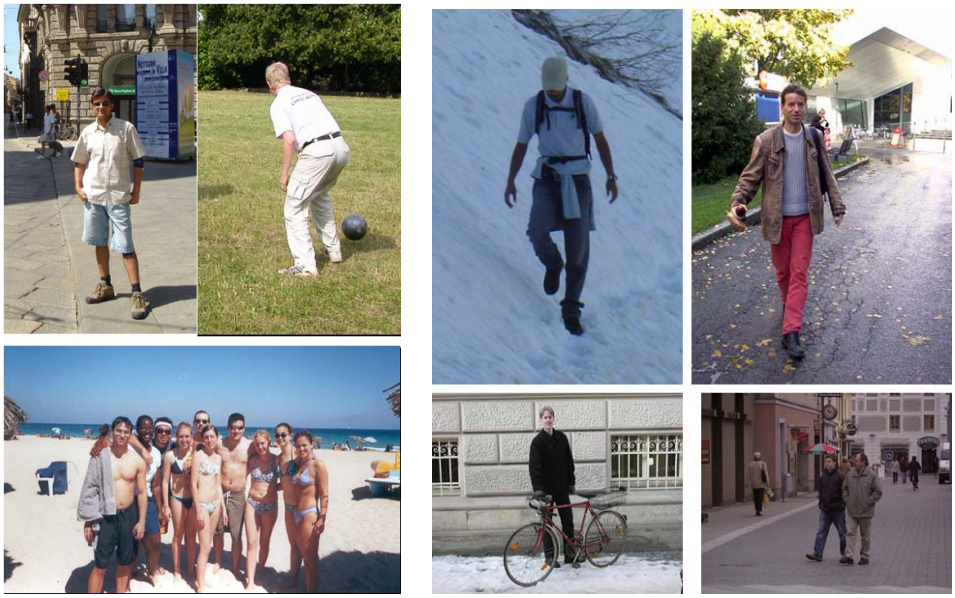

图 4-1　INRIA 中训练集（左）与测试集（右）数据

INRIA 数据集具有一些优点：数据图像较为清晰，图像中人体轮廓明确，图像背景多样。同时该数据集也有很多不足，例如图像中人体姿态单一，图像分辨率较低，数据集中训练与测试的数据量较少，对于更好的训练分类模型以及测试表征方法和分类器性能有所欠缺，与现实生活中实时监测数据还有一定距离。对于在自动驾驶、实时检测等应用方面，该数据集或许无法很好的模拟出现实场景下的数据图像。

（2） Caltech 数据集

Caltech 数据集是由加州理工学院拍摄制作，数据集中包括大约 10 个小时的 640*480 分辨率下 30Hz 的视频，拍摄环境主要为车辆在城市环境中行驶道路。在进行算法运算过程中，会从视频中提取 4250 幅图像作为训练样本，4024 幅图像作为测试样本。

Caltech 数据集图像均为车辆固定角度拍摄实时道路及周边状况图像，能够很好的模拟自动驾驶实用场景，道路上人体形态完整，姿态多样，然而该数据集中图像分辨率相对较低，仅有 640*480 像素大小，在车辆高速运行过程中，拍摄图像对人体描述性能较低，除此之外，数据视频是在美国拍摄，其道路环境与中国国内的道路环境相差较大，不足以满足国内道路交通视频的条件。

图 4-2　Caltech 中视频图像

（3） Pri-SDL 高分辨率数据集

为了更好的研究高分辨率数据对人体检测是否有帮助，同时针对国内真实道路交通环境做测试，我们拍摄了 60 分钟国内道路交通环境下高分辨率人体检测视频数据，图像分辨率为 1440*1080。

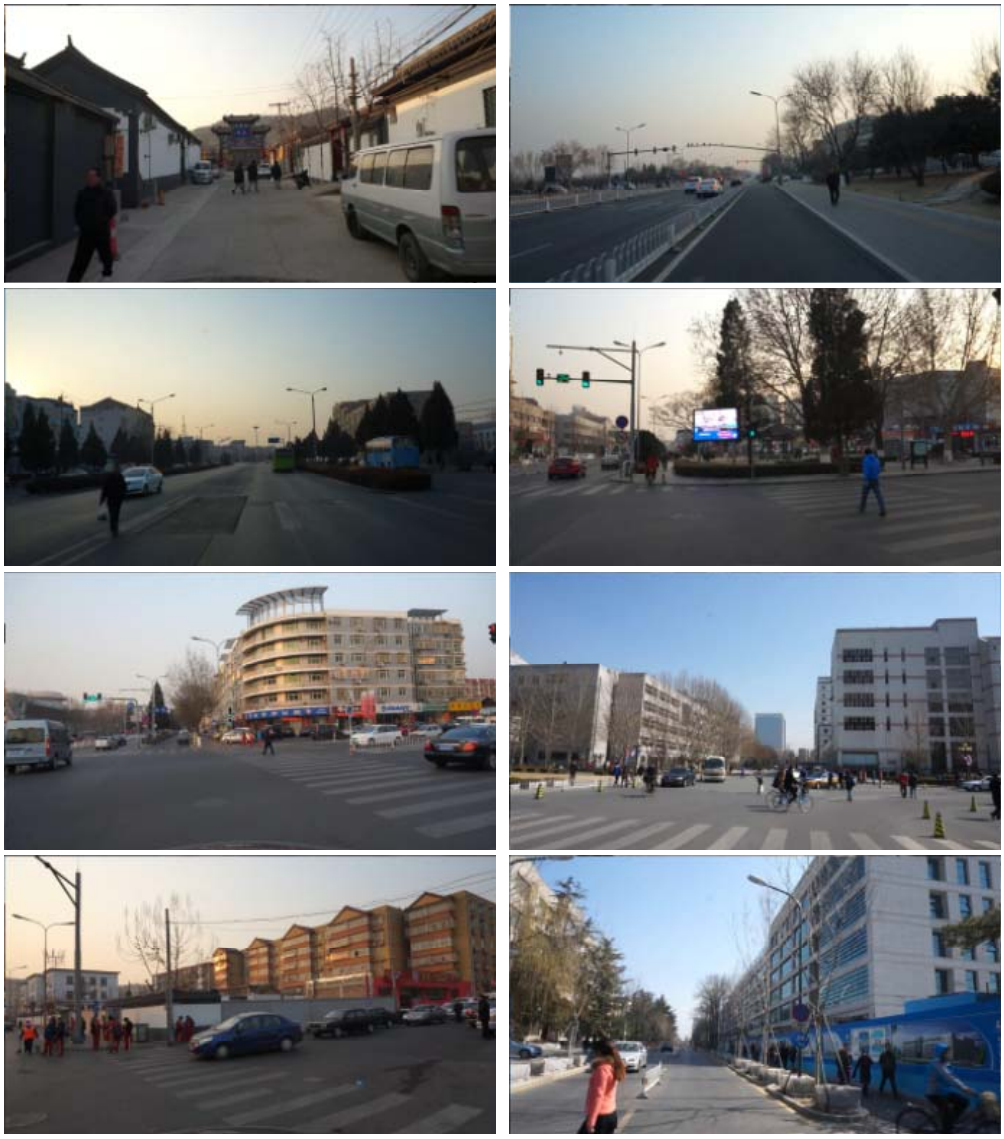

图 4-3　Pri-SDL 高分辨率人体数据集

研究数据拍摄的场景设计村庄、交通主干道路以及高校校园等，在保证场景多样化以及拍摄数据高分辨率的前提下，由于服饰以及光照会影响人体检测，因此还在不同季节拍摄不同时期的数据。同时，数据中人体数量比国外数据中人体数量增多，更能显示国内环境下真实的道路交通状况。

4.3　实验方法与对比结果分析

上一章介绍了 ACF 特征提取方法，以及基于多通道的稀疏重构特征提取方

法与卷积正交模板生成特征的方法，本文将采用这三种特征在高分辨率数据集上进行人体检测实验对比。由于我们将应用场景定位在自动交通驾驶方面，而在真实交通环境下，周围较远的人体检测是无意义的，因此在实验过程中，我们仅评测图像中人体像素高度超过 100 的人体检测效率。

我们提取了 4767 幅图像作为训练样本，其中标定了 7300 幅人体样本，提取 3177 幅图像作为检测样本，其中标定人体数量为 2500。利用上一章节介绍的三种特征提取方法在该数据集上进行实验，误检率如图 4-4 与图 4-5 所示。

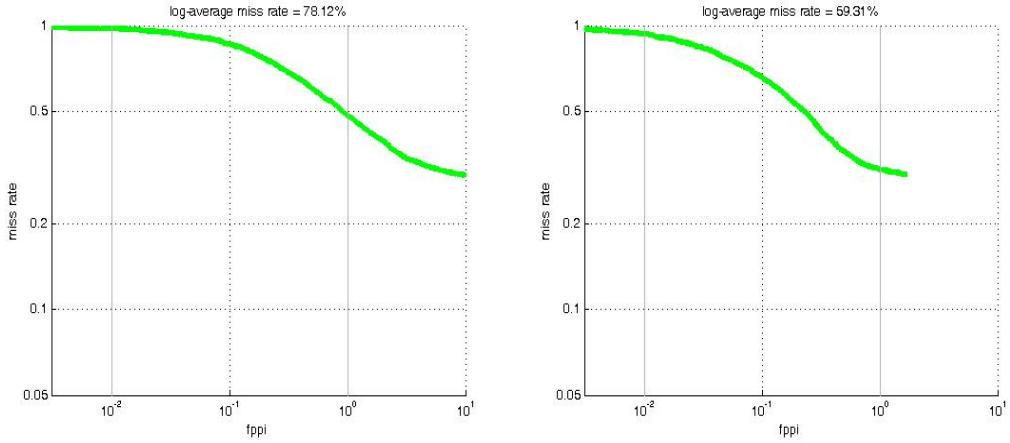

图 4-4　ACF 与重构特征在高分辨率数据集的性能曲线

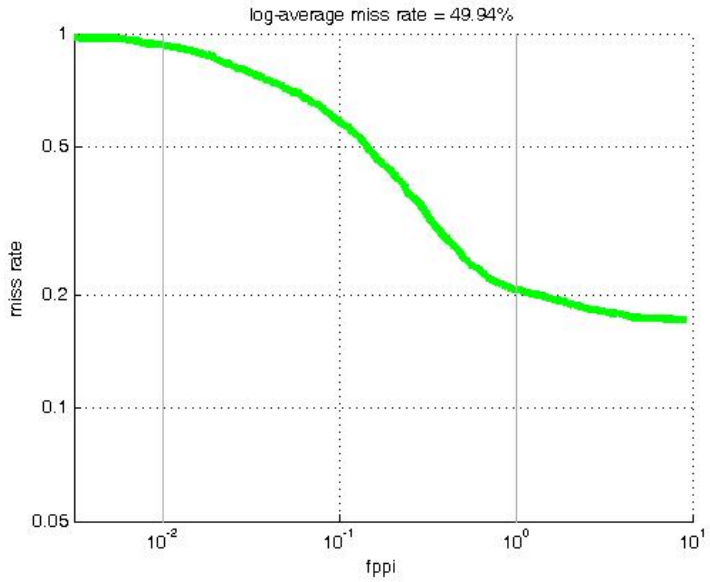

图 4-5　卷积正交模板特征在高分辨率数据集的性能曲线

通过结果图对比可见，本文在上一章节提出的两种特征在高分辨率数据上的性能比 ACF 特征提取方法的性能有了较大的提升，但是三种较好的特征在高分辨率数据上的检测结果距离实际应用的要求还有很大的距离。卷积正交模板生成特征的方法在高分辨率数据上的检测效果图如下：

图 4-6　检测结果示意图

在检测结果图看出，在多数人体检测上，结果较好，对于有少量遮挡的人体也能够检测，但同时也发现其中存在的一些问题。在图 4-7 中，红色框为未检测到的 groundtruth，黄色框为部分检测结果，在图 b 与 d 中，未检测到的人体衣服颜色与背景相似度较高，图 a、c 与 e 中的人体姿态与正面人体差距较大，且衣着颜色单一，说明在现实场景下，人体姿态与衣着颜色是影响人体检测的重要难点。

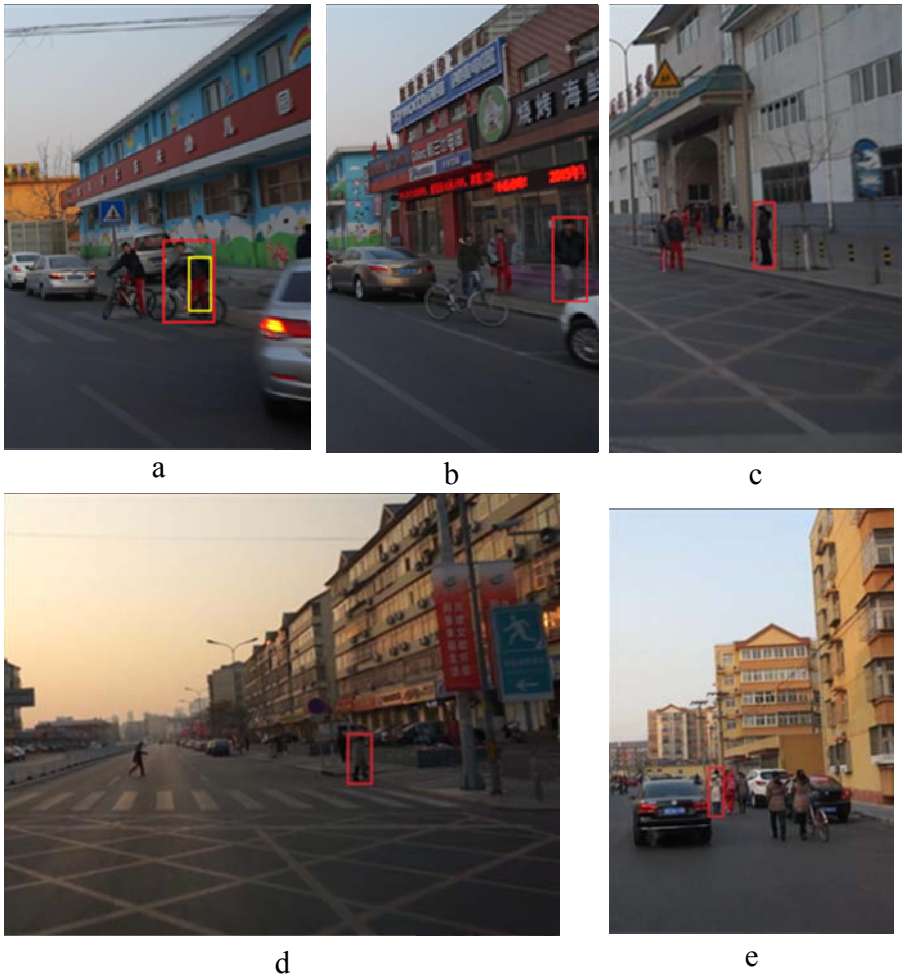

图 4-7　漏检图样

通过本节实验，我们发现高分辨率数据对于进一步提高人体检测效率具有一定的意义。在提取更加有效的特征描述子前提下，提高数据分辨率能够进一步增强人体检测的性能。

总结与展望

作为计算机视觉中目标检测的一个重要组成部分，人体检测在智能视频监控、智能人机交互等领域有巨大的研究价值与广阔前景，很多学者在这方面做了大量的研究工作，但由于面临多姿态、多视角以及遮挡光照等影响，一直未能广泛应用于实际生活。特征提取与分类器训练是人体检测框架中最重要的两个组成部分，本文针对人体检测中的特征提取这个问题，重点介绍了基于手工设计的特征和基于学习的特征两类提取方法，并分析两者的优势与不足，提出基于多通道的稀疏重构特征与卷积正交模板特征，在一定程度上提高了人体检测效率。除此之外，由于硬件设备的加强，已经很容易获得高分辨率的图像视频，随着人体检测陷入瓶颈期，本文试图探索使用高分辨率图像是否有助于提高检测效率，降低对特征提取与分类器训练的依赖性。

本文的主要工作如下：

（1） 基于人体检测实验框架中的常用的特征提取方法，从手工设计的特征与基于学习的特征两方面，研究了常用的人体特征，例如 HOG 特征、LBP 特征以及 CNN 特征等，并进行简单实验，分析各自的优势与不足。

（2） 重点研究结合两类特征的问题。基于 ACF 实验框架进行拓展，在多通道特征的基础上，结合稀疏模型，对不同通道的特征进行稀疏重构表达，从而降低了光照遮挡对图像本身的影响，最后生成一种较为复杂的特征描述子，提高了人体检测效率。

（3） 利用主成分分析法等对图像数据提取正交滤波模板，再用滤波模板与图像做卷积运算生成复杂的特征描述子，去除图像变量的相关性，减少冗余信息的表达。通过实验表明，正交模板的去相关性有助于提高特征表达能力，在一定程度上提高检测性能。

（4） 采集制作高分辨率图像视频数据集，并结合本文提出的几种特征进

行实验，结果表明在提取更加有效的特征描述子前提下，提高数据分辨率能够进一步增强人体检测的性能。

虽然本文在人体检测特征提取方面做了一些研究工作，但是由于时间原因与水平的限制，本文的研究工作仍然存在很多不足。其一，尽管本文在两类特征融合方面，获得了一些性能增长，但是并不代表任意两类特征进行融合都会提高检测性能。本文在其他方法的选择与分析方面还缺少更多的实验和研究。其二，在本文的实验框架中，虽然最后一步采用较为复杂的特征描述子能够提升性能，但是运算速率并不高，检测一幅图像，平均需要 2-5 秒的时间。其三，由于时间不足，对于高分辨率图像的标定并没有做到尽善尽美，在标定尺度上缺少分级。

针对以上的不足，在今后的工作中可能改进两个方面：

（1） 针对不同的两类特征，可以进行更多的实验，同时可以融合更多其他的信息来作为图像表征。

（2） 进一步完善高分辨率图像视频的采集和标定工作，分不同季节和时间段采集，标定可按照人体目标的大小、被遮挡的比例等方面分级标定并进行实验，希望从更多角度探索数据对检测效果的影响。

参考文献

[1] Dalal N., Triggs B.. Histograms of oriented gradients for human detection[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, 1: 886-893.

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

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

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

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

[6] Viola P., Jones M.J.. Robust real-time face detection[J]. *International Journal of Computer Vision*, 2004, 57(2): 137-154.

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

[8] Zhu Q., Yeh M.C., Cheng K.T., et al. Fast human detection using a cascade of histograms of oriented gradients[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2006, 2: 1491-1498.

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

[10] Mu Y., Yan S., Liu Y., et al. Discriminative local binary patterns for human detection in personal album[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2008: 1-8.

[11] Schwartz W.R., Kembhavi A., Harwood D., et al. Human detection using partial least squares analysis[C]. *In: Proceeding of IEEE International Conference on Computer Vision*, 2009: 24-31.

[12] Wu B., Nevatia R.. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors[C]. *In: Proceeding of IEEE International Conference on Computer Vision*, 2005, 1: 90-97.

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

[14] Nam W., Dollar P., Han J.H.. Local Decorrelation for Improved Pedestrian Detection[C]. *In: Advances in Neural Information Processing Systems*, 2014: 424-432.

[15] Girshick R., Donahue J., Darrell T., et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *In: Proceeding of IEEE Conference on Computer Vision and*

Pattern Recognition, 2014: 580-587.

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

[17] Jarrett K., Kavukcuoglu K., Ranzato M., et al. What is the best multi-stage architecture for object recognition?[C]. *In: Proceeding of IEEE International Conference on Computer Vision*, 2009: 2146-2153.

[18] 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.

[19] Sermanet P., Kavukcuoglu K., Chintala S., et al. Pedestrian detection with unsupervised multi-stage feature learning[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3626-3633.

[20] Ren X., Ramanan D.. Histograms of sparse codes for object detection[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3246-3253.

[21] Lim J.J., Zitnick C.L., Dollar P.. Sketch tokens: A learned mid-level representation for contour and object detection[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3158-3165.

[22] Zhang S., Bauckhage C., Cremers A.B.. Informed Haar-like features improve pedestrian detection[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 947-954.

[23] Van de Sande K.E.A., Uijlings J.R.R., Gevers T., et al. Segmentation as selective search for object recognition[C]. *In: Proceeding of IEEE International Conference on Computer Vision*, 2011: 1879-1886.

[24] Vincent P., Larochelle H., Bengio Y., et al. Extracting and composing robust features with denoising autoencoders[C]. *In: Proceedings of International Conference on Machine Learning, ACM*, 2008: 1096-1103.

[25] Salakhutdinov R., Hinton G.E.. Deep boltzman machines[C]. *In: Proceedings of International Conference on Artificial Intelligence and Statistics*, 2009: 448-455.

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

[27] 甘玲, 朱江, 苗东. 扩展 Haar 特征检测人眼的方法[J]. 电子科技大学学报, 2010, 39(2): 247-250.

[28] Lienhart R., Maydt J.. An extended set of haar-like features for rapid object detection[C]. *In: Proceeding of IEEE International Conference on Image Processing*, 2002, 1: I-900-I-903.

[29] 目标检测的图像特征提取之 HOG 特征[Z]. 2012-08-31. http://blog.csdn.net/zouxy09/article/details/7929348.

[30] Ojala T., Pietikäinen M., Harwood D.. A comparative study of texture measures with classification based on featured distributions[J]. *Pattern Recognition*, 1996, 29(1): 51-59.

[31] Ahonen T., Hadid A., Pietikainen M.. Face description with local binary patterns: Application

to face recognition[J]. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2006, 28(12): 2037-2041.

[32] Ahonen T., Hadid A., Pietikäinen M.. Face recognition with local binary patterns[M]. *In: Proceeding of IEEE Conference on European Conference on Computer Vision*, 2004: 469-481.

[33] 目标检测的图像特征提取之 LBP 特征. 2012-08-31. http://blog.csdn.net/zouxy09/article/details/7929531.

[34] Lowe D.G.. Object recognition from local scale-invariant features[C]. *In: Proceeding of IEEE International Conference on Computer Vision*, 1999, 2: 1150-1157.

[35] SIFT 特征提取分析. 2012-06-06. http://blog.csdn.net/abcjennifer/article/details/7639681.

[36] Deep Learning（深度学习）学习笔记整理系列. 2013-04-10. http://blog.csdn.net/zouxy09/article/details/8781543.

[37] Paisitkriangkrai S., Shen C., van den Hengel A.. Sharing features in multi-class boosting via group sparsity[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2128-2135.

[38] Bo L., Ren X., Fox D.. Multipath sparse coding using hierarchical matching pursuit[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 660-667.

[39] Dollar P., Wojek C., Schiele B., et al. Pedestrian detection: A benchmark[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 304-311.

[40] Vondrick C., Khosla A., Malisiewicz T., et al. Hoggles: Visualizing object detection features[C]. *In: Proceeding of IEEE International Conference on Computer Vision*, 2013: 1-8.

[41] Hariharan B., Zitnick C.L., Dollar P.. Detecting objects using deformation dictionaries[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 1995-2002.

[42] Zhou Y., Chang H., Barner K., et al. Classification of histology sections via multispectral convolutional sparse coding[C]. *In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 3081-3088.

[43] Chan T.H., Jia K., Gao S., et al. PCANet: A Simple Deep Learning Baseline for Image Classification?[J]. *arXiv preprint arXiv*:1404.3606, 2014.

[44] Ke W., Zhang Y., Wei P., Ye Q., Jiao J.. Pedestrian Detection via Principle Filters Based Convolutional Channel Features[C]. *In: Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015.

个人简历及论文发表

个人简介

姓名：张耀　　性别：男　　出生年月：1989 年 6 月　　民族：汉族

- 2008 年 9 月至 2012 年 6 月　天津大学　计算机科学与技术　学士
- 2012 年 9 月至 2015 年 6 月　中国科学院大学　工业工程　硕士

研究成果

- Ke W., **Zhang Y.**, Wei P., Ye Q., Jiao J.. Pedestrian Detection via Principle Filters Based Convolutional Channel Features[C]. *In: Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing* , 2015.（已录用）

致　谢

在中国科学院大学攻读硕士学位的三年时间，是我人生中一段非常重要的经历。我遇到了很多困难，也付出了努力，同时实验室良好的学术氛围、严谨的治学态度深深感染着我。在毕业论文完成之际，由衷的感谢这三年来曾经给与我巨大帮助的老师、同学以及亲人朋友。

首先，我要诚挚的感谢我的导师叶齐祥副教授。在我硕士期间，叶老师从生活和学业方面给予了无微不至的关心与指导。叶老师在学术科研上严谨求实，同时具有很强的洞察力与宽广的视野，是我深受感触。感谢焦建彬教授在学习和生活中对我的每一点指导，焦老师平易近人，待人温和，提醒我们要注重综合素质的提高。感谢韩振军副教授在理论学习和科研过程提供的帮助和支持。

其次，感谢模式识别与智能系统开发实验室中所有成员，感谢陈孝罡师兄、柯炜师兄在科研方面给与的帮助和指导，感谢李策师姐、高山师兄在生活中带来的种种帮助。感谢实验室的其他同学，我们一起共同学习，共同成长。大家在实验室一起渡过了三年快乐的日子，一起凝结成难忘的回忆。

感谢我的父母，感谢他们多年来一直给我最无私的帮助和鼓励，是他们让我坚持一直勇敢的向前走下去。感谢我的父母为我所做的一切，愿他们能够为我骄傲，愿他们永远健康。感谢我的朋友们，感谢张若男、张文韬、张宁、黎贺，感谢你们在三年来给我的支持和鼓励，在无助的时候给我力量。

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巨大的帮助。

最后，感谢参加论文评审和答辩的各位老师。

张耀

2015 年 5 月