监控视频的自动分析

(申请清华大学工程硕士专业学位论文)

培 养 单 位：计算机科学与技术系

工 程 领 域：计算机技术

申 请 人：张 启 龙

指 导 教 师：胡 事 民 教 授

二零一七年五月

监控视频的自动分析

张启龙

Abnormal Action Detection and Recognition for Examination Room

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Engineering

by

Zhang Qilong

(Computer Technology)

Thesis Supervisor：Professor Hu Shimin

May, 2017

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

随着社会安全需求的不断增长，监控摄像头的使用越来越平凡。 视频监控目前已经成为学校、车站、医院等公共场所最重要的安全防护措施。 作为智慧城市和智慧交通的重要组成部分，视频监控系统已经被广泛的应用到人们的生活中。视频监控的主要目的是监督人们在对应场所的一些不正常的行为。 因为这个目的，人的动作检测、跟踪和识别已经在计算机视觉领域被研究了很多年，有非常重要的地位。

本文主要研究针对考场的监控视频中人的异常行为的识别算法,主要分为 2 个部分，一个是老师的跟踪部分，另外一个是学生的动作识别部分，这 2 部分组成了考场异常检测系统。本文的主要工作包括有：

1)、提出了一种新型的基于检测的目标跟踪算法，该算法结合了可行变模型的目标检测和基于颜色信息的滤波跟踪算法，使得算法能够长时间的跟踪多目标，即使是在遮挡、跳帧的情况下也能够不丢失目标，具有较好的鲁棒性。

2)、提出改进了一种基于移动轮廓特征的动作识别算法，该特征在利用在视频的光流场上提取梯度直方图的方式，很好的将动作在时间上和空间上的信息都考虑了进去，达到了很好的效果，在本文的 1300 多个数据中识别率达到了 94.14%。

3)、设计了一种考场监控视频的自动分析系统，该系统通过对老师的轨迹绘制以及对学生的动作识别来进行异常行为的分析。

鉴于真实考场监控数据中异常动作较少的情况，本文建立了自己的考场监控视频数据集，该数据集包括了 9 种动作共 1300 多个视频，每一个动作都包含有至少 110 个小视频，该数据集能够比较好的反映考场的动作，并且拥有一定的多样性。本论文的所有算法均已经在数据集上测试，并且达到了比较好的效果。我们期望能够在更多的场景中测试我们的算法。

关键词：图像处理；目标跟踪；视频监控；动作识别

Abstract

With the growing social security needs, the use of surveillance cameras becomes more and more ordinary. Video surveillance has now become the school, station, hospital and other public places the most important security measure. As an important part of intelligent city and intelligent traffic, video surveillance system has been widely used in people's lives. The main purpose of video surveillance is to monitor people in the corresponding place of some abnormal behavior. Because of this purpose, human action detection, tracking and identification has been studied in the field of computer vision for many years and has a very important position.

This paper mainly studies the identification algorithm of human abnormal behavior in the monitoring video of the examination room. It mainly divided into two parts, one is the teacher's tracking part, the other is the student's action recognition part and these two parts compose the examination room abnormal detection system. The main work of this paper is as follows:

To begin with, we propose a new target detection algorithm based on detection. The algorithm combines the target detection of deformable part model and the filtering algorithm based on color information, so that the algorithm can track multi-target for a long time, even in the case of occlusion the case can not lose the target. The algorithm has good robustness.

Secondly, an improved motion recognition algorithm based on moving contour feature is proposed. The feature is used to extract the gradient histogram in the optical flow field of the video. It is very good to take the information of time and space in action. It achieves a very good effect in a more than 1,300 data and the recognition rate is 94.14%.

Thirdly, we design an automatic analysis system for the monitoring of video. The system analyzes the abnormal behavior of the teacher through the trajectory and the action recognition of the student.

In this paper we established our own examination room data, which contains more than 1300 video. Each action dataset contain more than 110 videos. It can reflect the real examination room well, which has a good diversity through a lot of people. All the methods have been tested on the dataset and the results are good. We would like to do

some more work on the different scene.

Key words: Image Processing; Object Tracking; Video Surveillance; Action Recognition

目 录

第 1 章 绪论

1.1 监控视频自动分析的研究背景和意义

随着社会的发展，人类的文明在不断在进步，人们的生活质量在不断的提升，然而，与此同时，各种各样的治安问题以及恐怖袭击事件也在世界各地不断的发生着，给这个社会添加了很多不稳定的因素，造成财产损失或者人员伤亡，威胁着人们的日常生活。因此，目前各个城市已经开始不断加速去完善社会治安，其中一个重要的手段就是增加城市的摄像头，加大城市的视频监控系统范围，进行智能监控。而且由于平安城市、智能交通、公平教育等政策措施的推广和深化，在商场、马路、教室、汽车火车站这些公共场所也有更多的监控需求。最近几十年来，视频监控系统已经开始逐渐被大量的应用在人们的日常生活中。甚至由于近年来硬件设备的不断更新使得监控摄像头的成本不断降低，很多家庭也开始安装智能摄像头来用于入侵检测以及老人监护等等。总的来说目前的全球监控行业正在以指数的形式增长，各种各样的安防企业以及从业人员在不断的涌出，目前已经逐步形成一个巨大的生态圈。

我国在安防行业的投入也是逐年扩大，在《中国安防行业“十三五”（2016-2020 年）发展规划》中指出，截止到 2015 年末，安防行业企业有 3 万余家，从业人员超过 150 万人，较 2010 年增长幅度均超过 20%。安防企业年总收入由 2010 年的 2350 亿元增加到 2015 年的 4900 亿元左右，增长了一倍，年均增长 15.8 %；安防行业年增加值由 2010 年的 850 亿元增加到 2015 年的 1600 亿元，年均增长 13.5%。并且到 2020 年，实现安防企业总收入达到 8000 亿元左右，年增长率达到 10%以上，行业增加值 2500 亿元。在地铁站、机场、商场、校园等公共场所内经常有成千上万的摄像头，并且在智能交通的政策支持下各个路口也遍地是摄像头。

面对这些摄像头，常规的方式是通过人力去进行 24 小时监控，一个或几个人坐在几十个屏幕面前进行人眼的实时监控，如果看到发生状况则及时报警，然而在目前中国人口逐渐老龄化的前提下，劳动力的价格将逐步上升，这无疑将会使得通过人为去监控视频成本上升。并且人为的监控需要考虑到人的客观因素比如劳累、生病等等，这些原因会导致监控效果的下降，容易出现误判。并且有科学研究表明，当人眼一直在相似的环境下关注后，很可能对视频中的物品产生忽略效应，而且极端情况下很有可能人的眼睛无法正确识别出需要找出的对象或者正在发生的事故。而且由于摄像头众多会导致产生的视频数据异常庞大，而通过传统的人眼监控将

无法进行索引，这样将大大减小视频监控的功能，许多场景都需要监控数据能够进行索引（比如识别罪犯的踪迹）。不能进行搜索将可能大大延误破案时间，导致罪犯逃脱。

因此监控视频的自动分析是十分有必要的，通过计算机视觉以及计算机图形学等相关技术的使用，可以将传统繁琐单调的人工检测转化为计算机自动的图像视频检测，帮助检测者进行监测工作，最后减少或者完全取代人力的劳动。而且，以往的传统人工检测手段往往是警察通过调取部分时间的视频来进行人眼的审查，需要耗费较长的时间，这种工作方式的效率是十分低下的，通过引用自动监控视频的自动分析系统，可以建立图像视频的相关索引，让监控系统可以自动在以往的视频中搜索需要的内容，利用现在计算机的超强计算能力，使得传统的监控人力得到解放,面对目前产生的如此海量的监控数据，也可以做到快速，精准的定位搜索。

在现实生活中有很多场景可以应用视频的自动分析系统，比如 1）在人流密集的公共场所对人员的行为进行自动的监控报警，防止有抢劫、踩踏、聚众打架和恐怖袭击等严重危害人身安全的事件，及时的发出警报，减小事件的损失；2）在公共交通场合对路口和信号灯的摄像头进行实时的分析，当有汽车有交通肇事的行为时能够及时的找到当时记录的视频以及相关的车辆和人物，并且可以在事故发生后即可报警做好后续的可能救援工作；3）在 ATM 取款机以及自助银行等安装视频的自动分析系统，对监控范围内的取款人员进行保护，如果出现抢劫等紧急状况能够第一时间报警，并且记录犯罪人员的脸部信息以及逃离的轨迹，帮助警方快速破获案件；4）在公共考场内安装监控视频的自动分析，对考试中的考生起到监督的作用，如果有疑似作弊的行为能够迅速精准的进行报警，以达到考试的公正公平；5）在家庭或者小区安装智能视频的自动分析系统，可以记录家庭人员的脸部信息，如果有外来不明人员的入侵，可以及时的反馈给主人，并且记录相关人员的信息，这样可以有效的保护人员的财产安全；6）在森林等易发生火灾或者急需要保护的地方安装监控视频的自动分析系统，通过自动识别火苗或者黑烟可以迅速的对火灾进行报警，及时的进行救援灭火，从而减小火灾所带来的损失。7）对于特定人员的轨迹搜索应用监控视频的自动分析，对于一些在逃的犯罪分子或者是走丢的孩子老人等等，通过已知的人脸特征和人体的生理特征，对需要的目标进行自动的查找，这样可以高效迅速的找到疑似人员，对于刑侦工作有着巨大的帮助。

由此可见，监控视频的自动分析对于社会的安全是有着重大的意义，人们的日常生活安全都可以被它所包含在内，是一张无形的安全带可以保障人们的生命财产安全。由于对社会有着如此重大的意义，在学术界这个领域一直是计算机视觉最热的方向之一，而且这个方向事实上需要结合到计算机视觉的很多其他方向的研

究比如目标的识别、目标的跟踪、以及异常行为检测等等。但由于监控摄像头往往会安放在各个不同的地方，而且经常是一些比较偏的地方，会有各个不同的场景，光照条件、遮挡、视角变化、尺度变化等等在目前并没有一个普适的方法。这个领域还需要人工只能、模式识别、图像处理等领域的学科交叉，是一个极具挑战性的领域。随着社会的发展，硬件的更新也在逐步的加快，原来国内监控视频一个很严重的问题就是摄像头的精度不高导致自动识别难度太大，现在随着智能城市和数字交通的一系列政策的支撑越来越多的老旧摄像头被换成新的高清的摄像头，这也是监控视频的自动分析成为可能的一个重要前提。我国目前的智能安防水平还处在初级阶段，离一些发达国家还有很大的一段距离，相关技术的研究和应用还不够成熟，相对于此领域对于社会的重要性，继续研究这一领域并将其实施到生活中显得非常有必要。监控视频的自动分析有着很多的应用场景，本文主要重点研究的是考场的监控视频的自动分析，本文所拿到的数据是河北省教育厅所提供的正规高等考试的数据，本文将主要对考场的监控视频自动分析算法进行研究。

1.2 监控视频自动分析的研究现状

监控视频的自动分析有着很广的应用场景，所以在计算机视觉已经研究多年，其中以欧美国家进入该领域最早。其中对于工业界而言，最早在 1997 年在美国由美国的国防高级研究项目总署成立了覆盖了多所高等院校组成的视频监控项目（VSAM），其中提出了他们是谁，他们什么时候动，他们在哪动以及她们在做什么等问题，并且开发实现了复杂背景下多运动目标的跟踪识别。然后 Wren 等在后来年研究出了个人搜索系统，系统通过在一台固定摄像机实现了一个在没有遮挡的情况下可以找到目标行人的系统。在 1998 年 Lipton 和其他人利用了多个摄像机来跟踪和检测目标物体，达到了比较好的效果。同时在欧盟等国家，比利时的 Katholieke 大学以及法国的国家计算机科学和控制研究院一起给国家机关研究了一些法庭、监狱等场景的自动监控系统。这些早期的视频监控自动分析项目都受限于早期设备的粗糙，摄像头分辨率不高以及镜头的畸变较大等。

然而随着半导体工业的不断发展，各种计算机和摄像头的价格不断降低，并且性能也在不断的提升，越来越多的企业开始进行监控视频方向的研究。像美国的 OnSSI(On-Net Surveillance Systems, Inc)、Object Video 以及中国的格林深瞳和 Face++还有商汤科技和海康威视等等。这些公司研究出了一些相当不错的自动监控系统以及一些算法，像 OnSSI 已经可以提供一套比较成熟的整体的视频监控系统，Face++的人脸识别技术已经运用到支付宝以及 APAC 会议的安全监控中，格林深瞳目前也已经设计好了一套完整的有自动分析功能的摄像机蓝图，这些公司

都获得了巨额的商业投资，在市场上取得了比较好的成功，然而，这些公司都还仅仅在初级阶段，良好的商业化盈利模式还并未找到。

监控视频的自动分析包含了许多，例如前背景分割，目标识别，目标跟踪，动作识别，异常检测甚至三维重建等等，在学术界，这些领域每年都有着大量的文章在国际上的重要期刊和会议（如 CVPR、ICCV、ECCV 等）上发表，一直是学术界最热门的课题。

1.2.1 前背景分割

前背景分割顾名思义就是将图像的前景和背景分离，而其中的依据是根据图像的一些属性比如颜色、灰度、纹理、和轮廓等，让那些属性不相同的部分能够表现出来，而在那些属性相同的部分可以划到一起，而前景部分则是我们感兴趣的部分，背景部分则是我们不感兴趣的部分。

人们经常在对图像和视频的研究应用中，往往都会仅对图像或视频的某些区域感兴趣，比如照片中的人脸、人行道上的行人或者考场中的固定考生座位等等。为了能够更好的分析这些感兴趣的位置的其他行为或者特征，就需要通过前背景分割来将它们分离开来，其中前景就是我们所感兴趣的区域。前背景分割技术是很多计算机视觉问题中的基本根基。

许多科学家都已经提出了自己的算法模型，Ridder[1]使用了信号处理上的卡曼滤波器来对前背景进行分割，其初衷是利用卡曼滤波器的动作预测能力，通过在前几帧的图像中对运动物体的运动估计，然后得到下一帧中前景可能存在的区域，这种估计问题在于如果一开始的目标就是错误的，那么后面的前景提取效果将会非常差，并且此方法使用的是运动的像素作为前景提取的条件，而前景提取的目的并非仅仅限于如此。Davis[2]通过使用一种运动模板的方法来找到视频中的运动区域，从而分离前背景。Elgamma[3]在其论文中使用了一种无参数估计的前景提取算法，Stauffer[4]等人在提取复杂的非静态前景的时候建立了高斯混合模型(MOG)，效果很好. Lee[5]将高斯混合模型和贝叶斯架构相联系。Harville[6]在高斯混合模型中加入了颜色和梯度的信息。Javed[7]在其中加入了均值转移算法。

前面的算法都考虑了比较复杂的情况，但是在质量和速度上都有待提高，并且当目标物体处在比较复杂的背景中时会很容易误判，并且边缘部分会很难区分，特别对于监控视频这样一个有实时需求的应用场景。

1.2.2 目标跟踪

监控视频的自动分析系统在识别出目标物体之后往往需要对物体的轨迹进行

刻画，通过目标跟踪可以完成这项工作。目标识别是在计算机视觉领域中一个重要的方向。高功率计算机的普及，高质量和便宜的摄像机的可用性以及对自动化视频分析的需求日益增加，已经引起了人们对对象跟踪算法的极大兴趣。

运动目标跟踪算法可以分为二类：生成模型方法和判别模型方法。其中生成类方法，在当前帧对目标区域建模，下一帧寻找与模型最相似的区域就是预测位置，而判别模型方法是图像特征加机器学习，当前帧以目标区域为正样本，背景区域为负样本，机器学习方法训练分类器，下一帧用训练好的分类器找最优区域。通常情况下，一个跟踪算法的好坏一般由其设计的和目标的相似度和特征直接相关，如果模板设计的好，那么跟踪效果将会很好，反之则差。而目标跟踪算法的速度一般由算法的搜索方式和滤波的方式所决定。

在生成类跟踪算法中，Kass[8]等人首先提出了 Snake 模型，此模型是一种基于能量最小化的跟踪算法，其设计了一种曲线，该曲线能够通过最小化能量函数来慢慢自动调整和目标物体的轮廓相一致。由于其能够自由变化，所以该模型能够跟踪任由形状的目标，其方式为先初始一个基本的边界然后通过公式描述目标的边界，最后通过逼近目标函数值的方法来进行迭代跟踪。Broida[9]等人提出了卡尔曼滤波跟踪器，其通过卡尔曼滤波器通过上一帧的物体的位置来预测下一帧物体的位置从而进行跟踪。Comaniciu[10]提出了一个无参数估计的均值漂移跟踪算法，它能够在一组数据的密度分布中寻找到局部极值，它比较稳定，而且是无参密度估计（它不需要事先知道样本数据的概率密度分布函数，完全依靠对样本点的计算），而且它在采样充分的情况下，一定会收敛，即可以对服从任意分布的数据进行密度估计。

在判别模型方法中，Hare[11]提出了一种新型的在线学习目标判别模型 Struck(Structured output tracking with kernels)，该算法通过避免传统的跟踪算法中选取了和目标不一样的物体作为分类器的训练样本，而是直接将输出空间控制在跟踪的目标内，这样就使得跟踪结果避免了训练分类器的误差，达到了很好的效果。并且该算法还在时间上通过设置阈值的方式控制分类器的过增长，让速度也得到提升。但是对于长时间的跟踪，Struck 算法仍然不能很好的完成。Kalal[12]等人提出了一个基于检测的跟踪算法 TLD(Tracking-learning-detection)，此算设计了一种可以不断在线学习的跟踪算法，通过不断在跟踪中跟新目标的特征点来不断重新训练检测器，从而达到跟踪能够解决遮挡、跳帧甚至丢失的情况，让跟踪显得非常鲁棒，然而在速度上其还比较慢。

1.2.3 动作识别

动作识别作为研究视频中人的行为相关的方向在计算机视觉领域有着重要的

意义，目前主要的识别方式分有二大类，第一类是基于模板匹配的方法，第二类是基于状态空间的方法。其中在第一类基于模板匹配的方法主要方式为先提取相关的有效动作特征，然后通过 bag-of-feature 等方法对特征进行重组，最后通过分类器来进行行为的识别。而基于状态空间的方法主要是把瞬间的姿态定义为一个状态，然后假设这些状态之间的转化是符合某种概率模型，然后进行建模通过训练队参数进行估计，然后再用模型进行预测。

Dollar[13]等人在 2005 年提出了一个基于时空特征的动作识别算法，他们展示了通常使用的 2D 感兴趣点检测器的直接 3D 对应物是不充分的，他们提出了一种 3D 的特征。针对这些兴趣点，他们设计了基于时空窗口数据的识别算法。Klaser[14]等人将传统的二维梯度特征扩展到了三维。Laptev[15]等人在 2008 年将光流作为一种特征引入了动作识别中产生了 HOF（Histogram of Optical Flow）特征。Wang H[16]等人提出了 HOG（Histogram of Gradiant）并在此基础上和 HOF 特征结合，提出了 MBH（Motion Boundary Histogram)特征，此特征结合了 HOF 和 HOG 二者的优点，将时间和空间都很好的考虑。其最后选择在运动轨迹中选取特征点来进行特征的提取，将 HOG、HOF、MBH 一起串联起来做成了 DT（Dense Trajectory）特征。成为了当时的 state-of-art。

前面的特征主要都是人工设计的特征，但这种特征由于是人工设计的，其泛化能力有限，基于深度学习和神经网络能够抽象出比较深层次的高级特征，在动作识别领域也有相关研究。Taylor[17]等人首次通过深度神经网络提取三维特征。Du T[18]等人提出了一种新型的 CNN 网络模型来提取视频中动作的三维特征。然而由于视频数据的多样性以及视频中人物的动作姿势的多样性导致目前在视频方向神经网络并没有像在图片领域取得如此大的成功。还有很多地方有待完善。

1.3 本文的主要内容

本文主要研究了监控视频的自动分析的课题，在介绍了目前的智能监控系统的工业界和学术界的发展状况之后，针对这个领域的一些个特定的问题和难点进行了探索和实验，其中主要对三个方面进行了研究，即视频中的前背景提取、目标跟踪以及动作的识别。

本文主要对考场这样一个特定的场景进行了视频监控的研究与优化，通过研究对老师的轨迹的跟踪以及同学的动作的识别，来对考场监控进行一个智能化的设计探索。其中在前背景提取方面主要借鉴了一种基于码本的动态阈值方法来提取动态的前景，在跟踪方面主要通过边检测边学习的方法来跟踪多个老师，在动作识别上设计了一个可以自动定位自动检测行为的方法来检测学生的行为，最后能够

对一些比较明显的异常动作进行识别并及时报警。

本文的主要工作有：

1) 设计了一种考场监控视频的自动分析框架，支持多种算法的灵活切换，包括跟踪算法、动作识别算法。
2) 研究前背景提取的算法，设计并实现了一种快速提取前景的图像算法并将其在目标检测中使用。
3) 研究动作识别算法，设计并实现了一种基于模板的动作识别算法，使用了一种自己优化过的特征。
4) 制作了一个考场场景的数据集，里面包括 6 个角度一共 4 个多小时的考场视频以及 3000 多个考生的动作视频样本。

1.4 论文的结构安排

论文共包含六章，其结构如下：

第一章， 绪论

主要介绍了论文的背景、题目的来源，通过对国内外自动监控视频领域的发展现况的研究，提出了本文所需要研究的内容。

第二章， 基于码本的前景检测算法研究

本章对以前的传统前景提取算法进行了总结和归纳，并且根据其所存在的问题进行了讨论和研究，提出了一种基于码本的前景检测算法，该算法首先通过记录码本来记录原始的静态颜色信息，然后不断通过动态更新来对运动物体进行定位，最后再通过形态学处理将以往漏掉的前景补齐，使得 recall 率大大提升。

第三章，基于动态识别的多目标跟踪算法研究

本章介绍了一些传统的目标跟踪算法，然后提出了一种基于检测的多目标跟踪算法，此算法通过目标检测来对跟踪的目标进行更新与矫正，这种方式不仅仅解决了遮挡、跳帧导致目标丢失的情况，还可以让跟踪框随着目标离摄像头的远近而进行变化而不需要进行摄像机的标定。

第四章，基于混合特征的动作识别研究

本章研究了传统的动作识别方法，并且对梯度、光流、轨迹、边缘、颜色等传统的特征进行了归纳，提出了一种改进的混合梯度光流边缘特征，这种特征很好的考虑了动作的时间性和空间性，在考虑动作之间在空间上轮廓的区别以外还加上了随时间变化而产生的位移的变化，这种特征可以很好的区别一些比较有作弊嫌疑的动作，对考场监控视频的自动分析有效。

第五章，考场自动监控系统框架

本章描述了考场监控系统的一个框架，该框架结合了前面的跟踪以及动作识别技术。然后介绍了一个真实的数据集，主要有 9 种需要的动作，一共有 4 个小时的视频以及 3000 多个剪裁好的动作小片段，由于一般而言考场的异常动作都会非常少见，所以很难找到有效的数据集来训练，本文建立了一个较大的数据集，很好的填补了这一方面的空白。

第六章，总结与展望

本章主要对本文所做的工作进行总结，并且对未来的工作进行展望。

第 2 章 基于码本的前景检测算法研究

2.1 引言

前背景提取方法是很多计算机视觉方向的基础，该问题研究的是如何通过简单有效的方法将图像中人们感兴趣的研究物体的部分提取出来，该有兴趣的部分就被称为前景，而背景则是人们所不需要或者不感兴趣的部分，而人们感兴趣的部分往往是那些运动的部分，通常情况下该问题的解决方法为通过对比与历史帧的区别来确定前景部分。在前景部分被提取完之后，就可以进行其他的一些计算机视觉领域的操作比如动作识别、目标检测等等。目前有许多背景减除算法，一般都是根据一些历史帧和当前帧的差异来进行提取，然而由于摄像机的效果受到光照等因素的影响强烈，因此有些算法只能适用于一些平缓的光照模型情况下才能有好的结果，并且由于后续的工作不同，对于前背景提取算法的要求也不同，有些前背景算法必须要移植到摄像机硬件上去，那么速度是其一个硬性要求；有一些需要在室外灯光照变化剧烈的地方进行使用，那么其对光照变化的考虑就需要更多一些。

2.2 前景提取方法的简介

一般而言，前景部分往往是运动的部分，而背景往往是静止的，前景提取的方法中基本上就是提取运动物体的部分。通常为了表达前景部分，都会使用掩膜（Mask）来进行前景部分的展示。

（1）帧差法

在图像处理过程中，帧间差分是通过计算相邻的帧的差值从而得到运动的区域的，通过图像的差值能够快速的检测出图像中运动目标运动的范围，因为其运动之后原来的位置的像素值将因为目标的移动而产生变化，这种变化能够使得相邻两帧的做差得到不一样的值。常用的帧差法有二帧差法和三帧差法。帧差法是通过由于物体运动会导致相邻几帧的图像中原来位置的像素值由于运动物体移动而发生变化，而没有运动的部分会保持原来的值基本不变，这样几帧相减，所得到的差值较大的地方就可以被认为是前景运动部分，此方法能够很好的运用在多个目标和摄像头移动的情况下，并且由于操作简单，其算法的速度可以非常的快，并且由于有阈值的选取所以对光照有一定的适应性，该算法为最常见的前背景分割方法。其通过 2 帧之间的差值来进行的操作在时域上相当于是做了一次高通滤波。该算法

的缺点也很明显，由于仅仅是进行相邻帧的差值来提取前景，如果物体运动过快而选取的帧间距离过大，那么很有可能同一个物体被分割成了 2 个物体；同样的，如果物体运动过慢而选取的帧间距离过小，那么由于物体还没有来的及变化导致减出来的图像没有物体。并且由于是相减，其只能提取前景部分的边缘，而不能将前景部分内部也划分出来，这对后面的应用操作也有一定的影响。二帧差法的具体过程如下：

选取图像中连续的二帧图像I_t、I_{t+1}，然后计算它们之间的差值，需要取得绝对值，如式（2-1）：

$$d_{i,\mathrm{j}}(x,y) = |I_{t+1}(x,y) - \ I_t(x,y)| \tag{2-1}$$

然后对得到的差值图像通过选择合适的阈值 T 来进行二值化从而产生掩膜，这个 T 值很关键，阈值表示如式（2-2）所示：

$$b_{i,\mathrm{j}}(x,y) = \begin{cases} 1 & d_{i,j}(x,y) \geq T \\ 0 & d_{i,j}(x,y) < T \end{cases} \tag{2-2}$$

最后得到的图像$b_{i,\mathrm{j}}(x,y)$中大于 1 的部分就是我们所需要得到的前景部分，然而这样的二帧差法会有问题，因为由于运动的物体运动过快，而一帧的间隔往往会比较慢（相对于运动而言），这样会使得提取的运动目标往往比真实的运动目标要大，通常就会出现“鬼影”现象。并且由于检测出来的物体是前后两帧的相对变化部分，无法检测到重叠的部分，这样会导致出现“空洞”。针对上述的 2 个问题，出现了三帧差法，即将相邻的三帧图像作为一组来进行差分，这样就能够较好的检测出运动目标的形状。

三帧差法：选取连续的三帧图像I_{t-1}、I_t、I_{t+1}然后进行两两差分：

$$\begin{cases} d_{i,\mathrm{j}-1}(x,y) \ = \ |I_t(x,y) - \ I_{t-1}(x,y)| \\ d_{i+1,\mathrm{j}}(x,y) \ = \ |I_{t+1}(x,y) - \ I_t(x,y)| \end{cases} \tag{2-3}$$

对得到的差值图像同时进行二值化得到$b_{i,j\ -1}\ (x,y)$、$b_{i+1,j}(x,y)$，然后在每一个像素点$(x,\ y)$将得到的二值图像进行逻辑与的操作，得到三帧图像中的中间帧的二值图像。

$$B_i(x,y) = \begin{cases} 1 & b_{i,j\ -1}(x,y) \ \cap \ b_{i+1,j}(x,y) \ \neq 1 \\ 0 & b_{i,j\ -1}(x,y) \ \cap \ b_{i+1,j}(x,y) \ = 1 \end{cases} \tag{2-4}$$

通过式（2-4）中得到的$B_i(x,y)$就是我们需要的掩码（Mask）。帧差法的基本算法框架结构如图 2.1 所示：

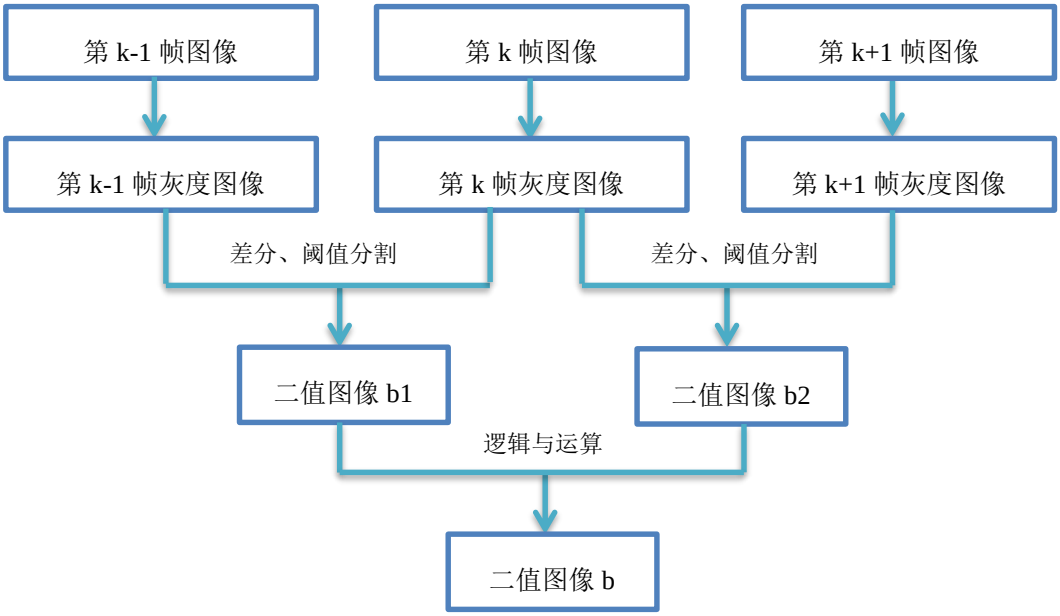

图 2.1　三帧法实现流程图

（2）背景减除法

背景减除法是一种基于数学建模的前背景提取方法，其通过一些数学的模型对背景信息进行建模，然后再对每个像素进行逐一的检测，如果该像素值的差别和该位置背景模型所预测的背景像素值相同或者在一定的阈值范围内，那么就认为此像素点为背景。事实上背景减除法可以看做是特殊的帧差法，其跟帧差法的区别在于帧差法是直接让背景相减，并没有对背景进行数学模型的建立，而背景减除法对背景进行了数学的刻画，这样的效果是使得该类算法往往有比较好的抗光照阴影的干扰能力，但速度会比简单的帧差法要慢。

背景差值法的前提假设是背景静止不变，即图像背景的像素值认为是不会变化的，其可表示为$b(x,y)$，我们定义图像序列为$f(x,y,i)$，其中(x,y)为位置坐标，i为图像帧数，将每一帧图像的像素值减去背景的像素值，通过式（2-5）这样就可得到一个差值图像：

$$d(x,y,i) = f(x,y,i) - b(x,y) \tag{2-5}$$

背景减除法检测前景由于进行了背景数学模型的建立，其准确性有了很大的提升相当于直接的帧差法，并且由于最后也仅仅是做了差值操作，所以速度也很快，然而由于要对背景模型进行建立，其检测效果受到模型的影响，如果模型建立的不好，那么其检测效果将受到影响，并且由于其建立的方式和光照有关，即对光照较敏感，并且在实际情况中背景往往是动态的，所以如何进行背景的重建是此方法的

一个关键点。一般是利用数学的手段提取有效的参数来建立一个背景的模型。常见的背景模型有：色彩空间码书法、单高斯分布模型、混合高斯分布模型、Kalman 滤波器法、HMM 模型法等。背景减除法其实可以看成是一种特殊的帧差法，只不过是直接减去背景图像。使用背景差分法需要对下面的一些问题进行考虑:

（一）背景获取：背景的数学模型参数需要背景的图像来进行训练，最好的情况是能够直接找到没有前景只有背景的图像，但如果遇到像城市道路那种一直有车辆经过的情况，只能找一个方法比如抠图等自行找到背景。

（二）背景的扰动：如树叶、树枝等各种东西的摇动。

（三）外界光照条件的变化。

（四）背景中固定对象的移动。

（五）背景的更新。

（六）阴影的影响。

本文 2.3 节将详细介绍一种基于色彩空间码书法的背景提取算法。

（3）光流分析法

光流的概念是 Gibson 等人在 1950 年首次提出来的。其是将现实三维空间上的速度的概念引申到了二维平面上，将空间上的物体在视频里所产生的在视频中的位移和帧的商来定义。空间上的物体的每一个点在图像上可以找到对应的像素点，空间物体的运动会导致图像上对应像素点的移动，那么这样就会产生一个运动矢量，即物体在图像上像素点的移动速度。一般情况下，光流是空间中前景本身在运动或者相机在运动或者二者同时运动所产生的。

如果视频中没有运动的物体，那么由于空间上的物体不动也不会导致图像上像素点的移动，那么光流场要么是静止的要么是相同方向的，但是如果有前景目标在移动，那么相对于背景而言其在图像上的像素点一定也在移动，这样求出来的光流场就会产生和背景不一样的方向或者大小，这样就能够提取运动的前景。然而光流场的计算比较复杂，需要进行方程的求解，这样对于前背景提取这样一个预处理的工作效率太低，一般达不到实时的要求，所以此种方法一般仅仅作为参考，目前还达不到实际运用的程度。

2.3 一种基于码本的背景减除法

简单的背景建模法[19][20]可以通过式（2-6）进行描述：

$$\mathrm{B_i}(\mathrm{i},\ \mathrm{j}) = \frac{1}{N}\sum_{t=0}^{N-1} X_t(i,j) \tag{2-6}$$

$$X_t(i,j) = \begin{cases} 1, & |I_t(i,j) - B_{t-1}(i,j)| > \mathrm{T} \\ 0, & |I_t(i,j) - B_{t-1}(i,j)| \le \mathrm{T} \end{cases} \tag{2-7}$$

其中式中的$B_i(i,\ j)$代表时刻 t 的时候的背景图像，$I_t(i,j)$代表的 t 时刻的当前帧，T 是选取的分割阈值，$X_t(i,j)$表示的是 t 时刻时得到的前景图像掩码。公式（2-6）就是均值背景模型的数学表达，当为中值模型时，背景模型即定义为 N 帧图像的中值即可，这种简单的背景模型好处是计算量非常小，算法速度特别快，然而缺点也很明显，由简单的均值、中位数、众数所得到的模型泛化能力特别小，不能够真正的反映出背景的像素值，背景图像的真实值与所建模型的真实值差别较大，所以不能够精确的提取到前景，而且此模型的背景是一开始就定义好了，并没有进行实时的跟新，然而在实际的视频背景图像是不断随着光照变化而变化的，随着时间的推移这种背景模型已经失效，所以需要更细致能够动态更新背景的模型才能真正运用到实际中。

然后有人就提出了基于混合高斯模型（MOG）来建立背景模型的方法[21]，其背景模型如下：假设每个像素用K个高斯函数描述，第 k 个高斯的权重为ω_k，那么在时刻 t，背景像素X_t的高斯模型可以描述为式（2-8）：

$$\mathrm{P}(\mathrm{X_t}) = \sum_{k=1}^{K} \omega_{\mathrm{k}} * \ \eta(\mathrm{X_t}, \mu_{\mathrm{k}}, \Sigma_k) \tag{2-8}$$

后续又有人将其他图像信息如颜色、梯度等加入到 MOG 模型中去，然而混合高斯模型由于需要提前计算出高斯参数，所以如果加上越来越多的图像信息，势必将导致计算量的增加，这样就必须在速度和准确度上进行一个权衡。

码本的前景提取基本思想是通过对背景像素的值建立一个码书来表示背景，相当于给背景像素通过一个压缩的形式来表达。这种思想类似于聚类算法。这允许该算法在有限的记忆下长时间捕获由于周期性运动引起的结构背景变化。相对于其他背景建模的算法，码本表示在使用空间和速度方面是有比较大的优势的，其避免了一些复杂的书序而模型的计算，没有参数的估计，想法也很直接。该算法的方法可以处理甚至包含有移动的背景或照明变化明显的场景，并且可以对不同类型的视频进行强大的检测。该算法的而且能够实时的更新背景的模型，并且没有任何对背景模型的先验知识，所以泛化能力特别的强，在检测的时候也是用的帧差的方法所以速度也非常的快，非常适合需要实时要求的一些场景如考场监控、交通监控等。

2.3.1 算法总体框架

Kim[23]等人提出的算法的基本数据结构有码本 CodeBook 和码元 CodeWord，对于图像的每一个像素，都将建立一个码本（CB）结构，每一个码本里面都将包

含有多个码元（CW）结构，其数学描述如下：

$$\xi = \{X_1、X_2、X_3、X_4 \ldots\ldots，X_N\}$$

$$\mathrm{CB} = \{\mathrm{CW}_1、\mathrm{CW}_2、\mathrm{CW}_3、\mathrm{CW}_4、\mathrm{CW}_5、\mathrm{CW}_6 \ldots\ldots \mathrm{CW}_n, \mathrm{L}\}$$

$$\mathrm{CW} = \{\mathrm{IHigh}, \mathrm{ILow}, \mathrm{max}, \mathrm{min}, \mathrm{t_{last}}, \mathrm{stale}, \mathrm{R}, \mathrm{G}, \mathrm{B}\}$$

其中ξ为训练背景的像素点集合，N 为像素的个数，L为一个CB中所包含的 CW 的数目，当 n 很小的时候，其退化为简单背景，当L较大时就可以对复杂背景进行建模;t 为CB更新的次数。CW是一个有 6 个维度的向量，其中 $IHigh$ 和 $ILow$ 是更新时的学习上下界，max 和 min 记录当前像素在学习中遇到的的最大值和最小值。上次更新的时间 t_last 和记录该 CW 多久未被访问的陈旧时间 $stale$ 用来进行判断 CodeWord 的使用情况，并删除掉比较少访问到的,后面的 R、G、B 为当前像素的 RGB 值。

Algorithm 1 CodeBook 算法

I.$L \leftarrow 0^1, \epsilon \leftarrow \emptyset$(空集)

II.提取需要的信息

for $t = 1 \rightarrow N$ **do**

(i) $X_t = (R, G, B), I \leftarrow \sqrt{R^2 + G^2 + B^2}$

(ii) 根据条件(a)和(b)在该像素码本的范围内找到和X_t匹配的码元CW_m

(a) $colordist(X_t, V_m) \leq \wp_1$

(b) brightness(I,$< IHigh, ILow >$) = true

(iii) If $\epsilon = \emptyset$或者没有匹配，then $L \leftarrow L+1$.创建一个新的码元CW_L并设置

•$\mathrm{CW}_L \leftarrow < IHigh, ILow, max, min, t, 0, R, G, B >$

(iV) 否则，更新匹配到的码元CW_m，假设其原有的属性为

$V_m = (R_m, G_m, B_m)$和$CW_m = (IHigh_m, ILow_m, max_m, min_m, tlast_m, stale_m)$更新

•$\mathrm{V}_m \leftarrow (\frac{min_m R_m + R}{min_m + I}, \frac{min_m G_m + G}{min_m + I}, \frac{min_m B_m + B}{min_m + I})$

•$\mathrm{CW}_m \quad \leftarrow < \quad min(I, ILow_m), max(I, IHigh_m), min_m + 1, max(max_m, t - tlast_m), stale_m, t >$

end for

III.对每一个码元$CW_i, i = 1, ..., L$,更新参数

图 2.2　codebook 算法伪代码

2.3.2 算法分析

算法对图像的每一个像素位置建立一个码本（codebook）,每一个码本有多个码

元（ce）多通道的，每一个码元都有自己的 *lowbound* 和 *upbound*。在学习时候，对于每一张图片中的每一个像素，进行相应的码本匹配，如果像素值在码本中某码元的 *bound* 范围之内，那么只需要更新该码元的 *bound* 即可。如果在对应的码本中没有匹配的码元，那么证明背景是动态的，需要在此像素点的码本中建立新的码元。因此，在背景学习的过程中，每个像素有多个码元进行对应，这样学习出来的模型就可以刻画动态背景。

相对于以往的混合高斯模型，这里的背景建模并没有任何数学参数，即并没有假定其符合某种数学概率模型，这样的好处在于并不是所有的前背景分布都符合高斯模型，而高斯模型也仅仅是可能最符合前背景分布的数学模型，所以此方法完全放弃了数学的模型，完全从图像本身和前背景本身的区别入手，定义了一套新的颜色空间。

为了处理全局和局部照明变化（如阴影和高光），算法通常采用标准化颜色（颜色比）。这些技术通常在图像的黑暗区域很差。暗像素比亮像素具有更高的不确定度，因为颜色比不确定性与亮度有关。亮度应作为比较色彩比例的因素。这种不确定性使得黑暗区域的检测不稳定。错误的检测往往集中在黑暗的地区周围。这个问题在[22]中有所讨论。因此，由于观察到像素值在照明变化下随时间的变化，照明通过减少或增加光强度随时间而变化，以使像素值更暗或更亮。像素值主要沿着朝向原点的轴分布成细长的形状，根据这个观察，算法设计了一种新的颜色模型，将颜色和光照信息的变化进行了一个很好的记录，当输入为像素 $X_t = (R, G, B)$ 以及一个码元为 CW_i, 其记录的 $RGB = (R_i, G_i, B_i)$ 时：

$$\begin{aligned} \left||X_t|\right|^2 &= R^2 + G^2 + B^2 \\ \left||V_t|\right|^2 &= R_i^{\,2} + , G_i^{\,2} + B_i^{\,2} \\ < X_i, V_i >^2 &= (R_i R + G_i G + B_i B)^2 \end{aligned} \tag{2-9}$$

那么颜色的失真可以描述为：

$$p^2 = \left||X_t|\right|^2 \cos\theta^2 = \frac{<X_i, V_i>^2}{\left||V_i|\right|^2} \tag{2-10}$$

$$\text{colordist}(X_i, V_i) = \sigma = \sqrt{\left||X_t|\right|^2 - p^2} \tag{2-11}$$

这种颜色失真的描述可以称为带有光照权重的颜色空间，这相当于将码字矢量几何重新缩放（归一化）为输入像素的亮度。这样，为了测量颜色失真，考虑到亮度，避免了标准化颜色的不稳定性。

算法在检测的时候还考虑到了光照的变化，通过存储 *IHigh* 和 *ILow* 这 2 个在全局码元中最大和最小的亮度值，可以控制光照的变化在一个合理的范围内这样

使得阴影和高光的影响都降低到最小。其中这个范围$[IHigh, ILow]$定义为：

$$\text{ILow} = \ \alpha * \min, \text{IHigh} = \min\{\beta * \max, \frac{min}{\alpha}\} \tag{2-12}$$

在这里α和β都是参数，其中α的范围在 0.4-0.7[3]，β在1.1-1.5[4]。而且这个限制范围在算法不断迭代更新码本的时候会收敛到一个定值，而在原来定义的光照逻辑公式如（2-13）所示：

$$\text{brightness}\big(\text{I}, (\max, \text{low})\big) = \begin{cases} true & \text{if ILow} \le \left|\left|X_t\right|\right| \le IHigh \\ false & \text{else} \end{cases} \tag{2-13}$$

通过上面的式子可以很好的将背景中光照的影响所区别，这样能够更好的找到前景目标而不被阴影或者高光所影响。由于其每个码本拥有多个码元，所以也可以很好的将多样化的背景所剔除。

2.3.3 算法改进

该算法由训练和检测两个部分组成，其中训练是假定在前景部分暂时没有出现在像素中时的时候进行，然而由于全局光照的影响，使得背景很难一直静止，它们会引起过度检测，误报以及对真实目标的不敏感。所以这就需要有一定的动态自适应措施，由于前景的加入导致前背景的区别，我们可以将新的图像也加入到训练过程中，即训练不仅仅只在初始化的时候进行，而是间断的进行。将训练集V_m进行动态跟新可以得到式（2-14）：

$$V_m \leftarrow \gamma X_t + (1 - \ \gamma)V_m \tag{2-14}$$

然后在后面继续添加式（2-15）的变化值：

$$\sigma_m{}^2 \ \leftarrow p\sigma^2 + (1 - \text{p})\sigma_m{}^2 \tag{2-15}$$

然后再运用到伪代码的步骤 II （iv）步进行迭代。其中$\sigma_m{}^2$是一个全局颜色失真参数，这个参数是针对新的颜色空间，而并不是传统的 RGB 空间，σ_m将在算法开始的时候初始化，最后颜色距离公式（2-10）将要变成如式（2-16）：

$$\text{colordist}(X_i, V_i) = \frac{\sigma}{\sigma_i} \tag{2-16}$$

考虑到码本有多个并且肯定有冗余，所以当训练结束之后需要对长期未访问的码元进行删除，设置阈值时间一般为学习时间的一半。这样处理之后能够节省很大一部分内存空间，使得算法再运行的时候搜索过程变得迅速，并且准确率保持原来的水平。

2.3.4 实验结果

本次的实验数据为真是的考场监控视频，其分辨率为 720*480，实验所使用的机器为英特尔酷睿 core 2.00GHz CPU 内存 32 GB RAM 和一块 GTX 980 GPU.训练帧数选取为 10 帧，最大迭代学习次数为 25，阈值 T = 30，跟新间隔为 50 帧，视频的主要场景为教室，里面有桌子椅子以及会走动的 2 个老师，其中由于摄像头的清晰度非常的低，是老旧的摄像头，所以会出现光照严重抖动的情况，并且还会有阴影和滞留的像素变化。在这种情况下一般的算法如帧差等等效果会非常不好，并且容易出现抖动的现象，由于光照的突然变化会很容易被认为是物体的突然移动。实验的数据为真实的考场监控视频数据，由河北省教育厅所提供，该视频一共有 3 段，其中在本文中用来做前背景检测的仅有当前一段视频。

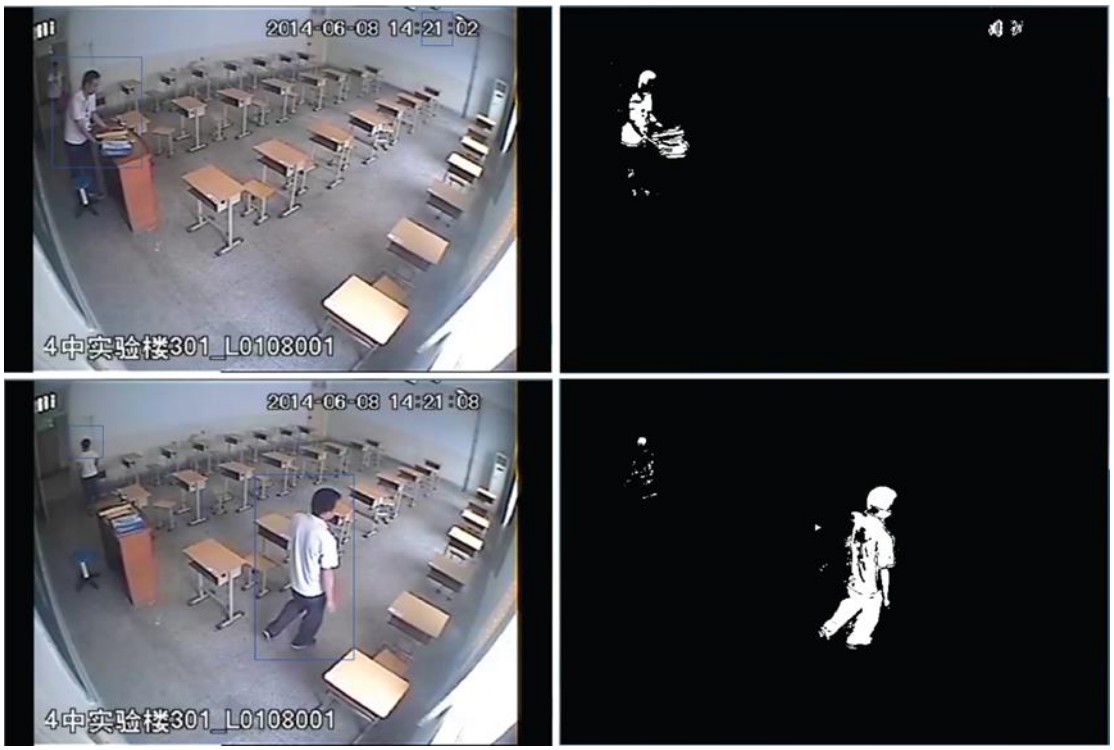

图 2.3 codebook 算法在真实教室里的效果

可以看到图 2.3 中所示结果显示该算法效果较为鲁棒，除了在走动的前景部分（老师）以外其他地方基本上都背认为是背景，除了右上角由于时间的变化也可以认为是前景。但是有一个问题，可以看到第二张图片中当老师停止下来之后其也会被该算法渐渐忽略，由于此算法是前背景提取算法，所以此现象是可以接受的，在目标跟踪的时候才会考虑其静止时候也要跟踪的情况。仔细看上图还可以发现此算法提取的结果不仅有前景的轮廓信息，而且还有内部的信息，完全可以通过肉眼

就看出来其为一个行人，这对于后续的形态学提取整个老师也很方便，基本上不用什么很复杂的操作就很容易能够将老师整个覆盖住。

图 2.4 背景减除的实际效果图

可以看到第一行的图中讲桌因为人的阴影改变了讲桌的像素值所以被认为是前景（运动了），然后阴影走过去之后其又被认为是背景。算法的 FPS 可以达到 30，并且鲁棒性较好，这个前景算法为后面的目标跟踪有着很大的帮助。但是由于是动态更新的，所以导致如果说目标静止过久其也会被认为是背景，但这种情况对于前背景提取算法而言是正确的情况。下面这个实验结果也可以看到其前景提取的准确度较高，除了目标之外没有其他物体被提取出来，而且对于目标的整个轮廓和内容都有提取到，而且其内容的提取比较完整，这对于后面的目标检测是有很大的帮助的，后续再通过形态学和连通域的处理就能够很好的将目标框出来。

虽然在开始的提取的时候需要有一定的训练样本来对码本进行训练，但由于考场这个场景是存在这样的情况的，所以这个算法针对考场而言非常的适合这个场景，算法在前景移动的时候能够有效的检测到其相对于当前位置和移动位置的像素值的差距，并且通过阈值的方式将该差距控制在一定的范围内，从而有效的解决了阴影使得差值变化过大的问题，在进行码本存储的时候使用的是数组结构，使得该算法在查找的时候非常迅速，这也是本算法在相对混合高斯模型等复杂模型方法上的优势所在。

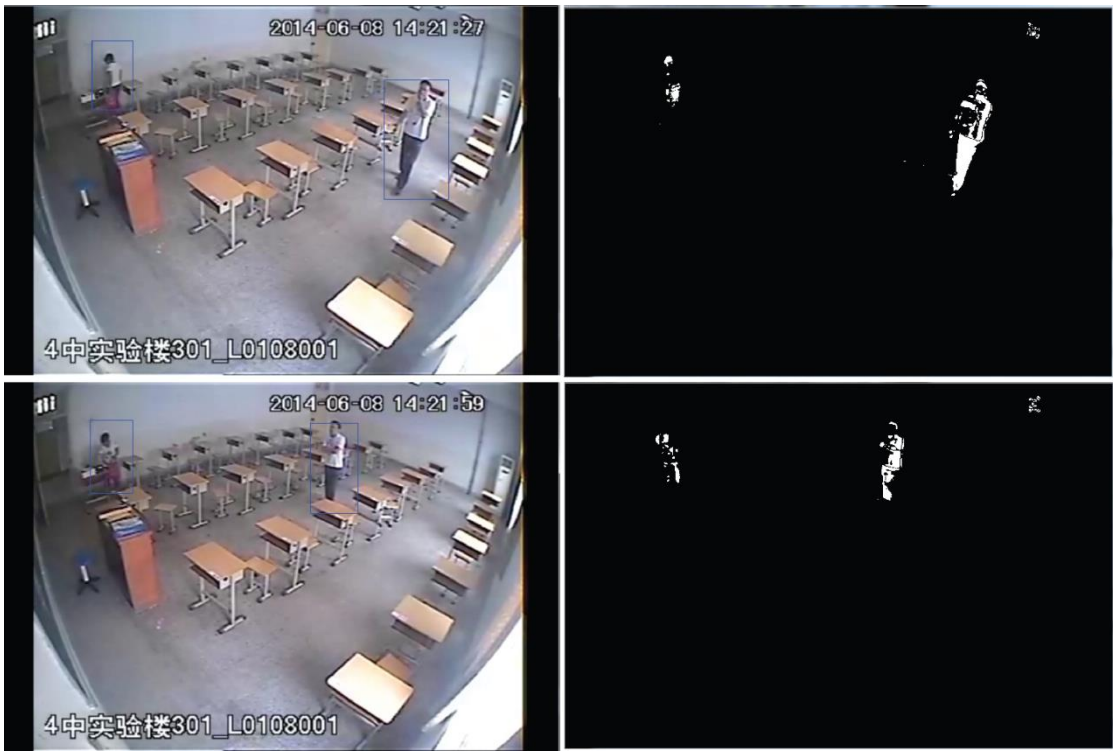

图 2.5 后续实验结果图

2.4 本章小结

本章首先介绍了前背景提取算法的功能以及目的，然后又介绍了目前几种传统的视频前景提取算法，分析了它们之间的联系和区别，优点和不足。接着主要介绍了基于背景减除的方法，并且引入了基于码本的背景减除方法，此方法由于并不需要先验假设数学模型，而是直接通过类似分类的方法将前景和背景进行记录区别，有效的避免了前背景不符合先验假设等情况，经过试验证明此方法确实有效，能够很好的完成考场视频监控的前景提取工作。

第 3 章 基于动态识别的多目标跟踪算法研究

3.1 引言

目标跟踪一直是计算机视觉领域的热门方向，其应用场景非常广泛，可以说只要和视频分析有关的应用和它都离不开，比如自动驾驶，导弹巡航、防空系统、无人机驾驶、VR 交互等等。视频目标跟踪的目的在于对每一帧图像，都能够根据一定的跟踪算法找到所需要得到的目标，最后将所有帧对该目标的位置刻画连接起来刻画出其轨迹，即将以个视频中的不同物体进行分类，并且根据某种匹配算法定位感兴趣的目标，并且还要完整的画出目标区域的位置。目前，目标跟踪技术主要应用于以下领域：

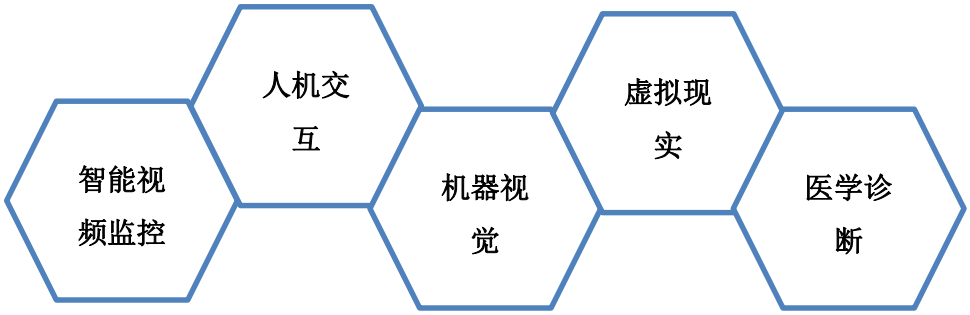

图 3.1 目标跟踪技术主要应用方向

1).智能视频监控：基于特征的目标跟踪，自动轨迹刻画，越界检测以及交通事故检测等等；

2).人机交互：虚拟现实产品的诞生导致原来的电脑的交互方式已经不够，为了使计算机具有识别和理解人的姿态、动作、手势等能力，跟踪技术是关键；

3).机器人视觉导航：在智能机器人中，跟踪技术可用于计算拍摄物体的运动轨迹；

4).虚拟现实：在虚拟的三维场景中目标跟踪才能够准确的定位人体的移动和动作的识别，目标跟踪使得该应用能够顺利的与人进行交互，手势识别、步态识别等都需要目标跟踪来进行铺垫；

5).医学诊断：跟踪技术主要应用在超声波和核磁序列图像，因为前面几种医学图像的噪声非常严重导致信息无法正确在单帧图像中获取有用信息，而目标跟踪技术使得可以在时间上和几何上找到这些医学图像中的有用信息，使得这些医学

检测结果更加准确。

在所有的跟踪算法当中，Kalal[12]等人提出了一个基于检测的跟踪算法 TLD（Tracking-learning-detection）是最适合长时间跟踪的一种算法，本文的多目标跟踪算法的思路取自于他。

本章的具体结构安排为：3.2 节介绍 TLD 方法，3.3 节主要介绍本文提出的一种新的多目标跟踪的算法，3.4 节对真实的考场数据进行试验结果的展示以及分析，3.5 节为总结了本文算法的优点和不足。

3.2 基于检测的跟踪算法介绍

传统的检测算法一般是通过目标检测或者人工的方式找到所需要跟踪的目标，然后再将目标进行特征的提取或者模板的提取，最后引入跟踪算法进行跟踪，而后续仅仅是跟踪模块的运行，这种算法有一种致命的缺陷，就是如果进行长时间的跟踪，目标物体不可避免的会发生尺度变化、被遮挡、光照变化或者形状发生变化等等，在这种情况下，现有的单纯的跟踪算法都有可能将目标丢失，由于没有有效的更新和纠错机制，导致目标丢失之后就难以再继续跟踪，这样就导致了在长时间跟踪领域并没有好的算法来实现。因此，对于长时间的跟踪，最重要的问题就是要找到一种能够不断对目标进行更新能够不断纠错学习的跟踪算法，这种算法要在目标产生尺度变化、形状变化、被遮挡的情况下能够对模板和特征进行更新，让跟踪算法能够继续识别已经发生变化的目标物体。在这种情况下，Kalal[12]等人提出了一个基于检测的跟踪算法框架，其结构如下：

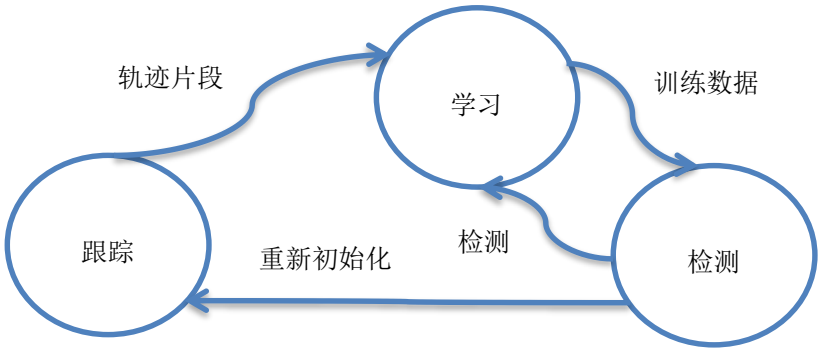

图 3.2 tracking-by-detection 算法结构

可以看到该算法框架主要由三个模块构成：追踪器(tracker)，检测器（detector）和学习器（learning）。

一）追踪器：这里的追踪器就是传统的跟踪算法中，仅仅在目标一直可见并且

没有产生巨大的形变或者光照变化时才有效，该算法可以准确的跟踪并形成一条目标的轨迹，其中这些轨迹上的目标都可以被认为是目标的正样本而拿到学习模块中去学习（Tracking->Learning）。

二）检测器：检测器是判定当前跟踪器是否为好的跟踪器的标准，其对每一帧都进行全面的扫描，并且找到目标和类似目标，将它们当做正样本和负样本，然后交给学习器来进行训练（Detection->Learning）。检测器还会选出一个当前最相似目标物体的帧作为当前的输出，并且更新跟踪器的起始位置，从此位置重新开始跟踪（Detection->Tracking）。

三）学习器：学习器不断通过检测器和追踪器给予的样本进行学习，然后改善检测器的精度（Learning->Detection）。

Kalal[12]等人提出的检测器和追踪器可以进行改进，根据其思想，3.3 节将介绍本文的检测跟踪算法。

3.3 基于梯度检测和颜色跟踪的检测跟踪算法

本文的 3.2 节提到了通过检测来跟踪能够对于长时间的跟踪有很高的鲁棒性，相对于其他的跟踪算法，其能够比较鲁棒的对目标进行长时间的跟踪，受到其思路的启发，对于考场监控视频我们提出了自己的跟踪算法，该算法利用 Felzenszwalb[24]等人提出的 Discriminatively Trained Part-Based Models（DPM）的模型进行目标的检测，然后通过 Danelljan[25]等人提出的自适应颜色命名跟踪器进行跟踪。

3.3.1 可变形多部分模型检测（DPM）

此模型是基于梯度直方图(HOG)[26]做的改进，其首先对目标进行 HOG 特征的提取，只不过将目标并不是整体提特征，而是将目标的部分都进行了特征的提取，然后认为模型是由若干个部件所组成，在检测的时候就利用多部件的权重加权来进行目标的检测，定位是通过移动窗口来实现的。

（1）HOG 特征

HOG 特征是一种基于图像梯度的特征，其基于的主要思想是图像的边缘的密度方向和大小能够很好的表现图像的形状，这样在检测的时候就能够被区别开来，其做法首先是将图像分成一个个小的细胞单元，然后再对细胞单元提取目标的梯度的方向与大小，然后根据方向的度数进行直方图的统计，最后再将各个胞元的直方图向量全部串联起来就成为该图像的 HOG 特征。由于该特征的提取是基于细小的胞元的，所以其对几何和光学都有很好的不变性。HOG 特征在提取的时候为了

保证更好的区分度，其在一个比胞元更大的环境下进行提取了归一化（一般为 4 个胞元），这样其对光照和阴影的效果又会有很大的提升。

HOG 描述子在图像的目标检测领域有着很好的效果，特别是对行人检测，其特征经过归一化后显示的效果基本上在肉眼上就能够判断其是一个行人。即使是出现一些细微动作和姿势，其也能够很好的区别。并且由于其提取的特征是基于直方图的特征，所以计算起来也很方便。

（2）DPM 特征

DPM 首先采用的是 HOG 进行特征的提取，但是又有别于 HOG，DPM 中，只保留了 HOG 中的 Cell。假设，一个 8*8 的 Cell，将该细胞单元与其对角线临域的 4 个细胞单元做归一化操作。提取有符号的 HOG 梯度，0-360 度将产生 18 个梯度向量，提取无符号的 HOG 梯度，0-180 度将产生 9 个梯度向量。因此，一个 8*8 的细胞单元将会产生，(18+9)*4=108，维度有点高，DPM 给出了其优化思想。首先，只提取无符号的 HOG 梯度，将会产生 4*9=36 维特征，将其看成一个 4*9 的矩阵，分别将行和列分别相加，最终将生成 4+9=13 个特征向量，为了进一步提高精度，将提取的 18 维有符号的梯度特征也加进来，这样，一共产生 13+18=31 维梯度特征，如图 3.2 所示。

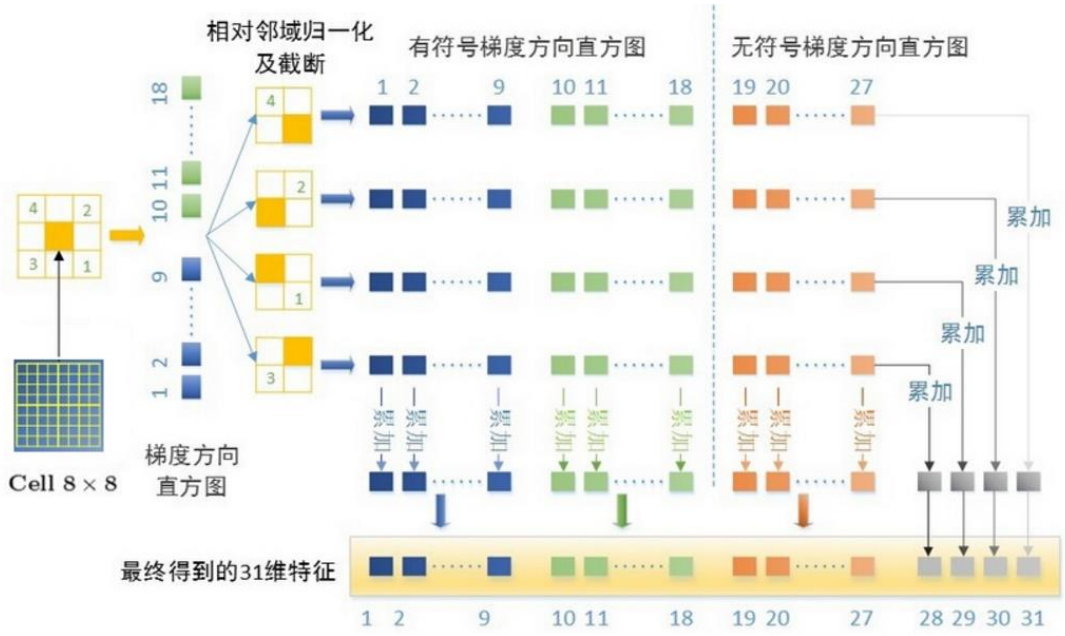

图 3.2 DPM 特征构成图，首先通过窗口在图像像滑动，每个窗口有 4 个块，每一个块又由 4 个胞元所组成，对每个胞元提取梯度直方图之后再进行累加，最后串联成一个特征向量。

（3）高斯金子塔

由于需要对目标进行部件的识别，如果原始图像的目标太小那么很难将部件准确的区分，这里就需要使用到图像金字塔，在这里高斯金字塔的尺度可以和 lowe[27] 在 SIFT 中的一致。

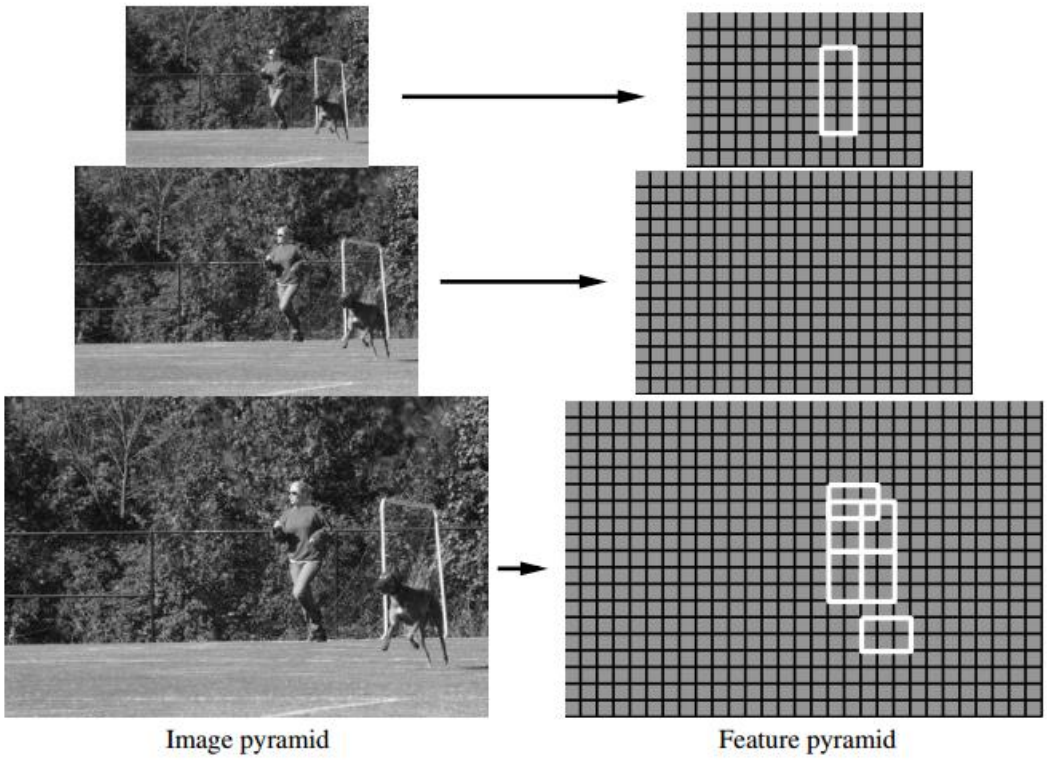

图 3.3 DPM 中的高斯金字塔[13]

（4）部件响应

再得到各个部分的特征之后需要对模型的部件进行响应，这样利用部件的响应值的投票来综合打分，进而对目标进行判别，其中响应值的公式（3-1）：

$$\text{score}(x_0, y_0, l_0) = R_{0,l_0}(x_0, y_0) + \sum_{i=1}^{n} D_{i,l_0-\lambda}(2(x_0, y_0) + v_i) + b \qquad (3\text{-}1)$$

其中，x_0, y_0, l_0分别为锚点的横坐标，纵坐标，尺度。$R_{0,l_0}(x_0, y_0)$为根模型的响应分数，$D_{i,l_0-\lambda}(2(x_0, y_0) + v_i)$为部件模型的响应分数，$b$ 为不同模型组件之间的偏移系数加上这个偏移量使其与跟模型进行对齐，$2(x_0, y_0)$表示组件模型的像素为原始的 2 倍，所以，锚点*2，v_i为锚点和理想检测点之间的偏移系数，部件模型的详细响应得分如式（3-2）：

$$D_{i,l}(x, y) = max_{dx,dy}(R_{i,l}(\mathrm{x} + \mathrm{dx}, \mathrm{y} + \mathrm{dy}) - d_i * \phi_d(dx, dy)) \qquad (3\text{-}2)$$

其中，x, y 为训练的理想模型的位置,$R_{i,l}(\mathrm{x} + \mathrm{dx}, \mathrm{y} + \mathrm{dy})$为组件模型的匹配得

分,$d_i * \phi_d(dx, dy)$为组件的偏移损失得分,d_i为偏移损失系数,$\phi_d(dx, dy)$为组件模型的锚点和组件模型的检测点之间的距离。

简单的说，这个公式表明，组件模型的响应越高，各个组件和其相应的锚点距离越小，则响应分数越高，越有可能是待检测的物体。

3.3.2 基于自适应颜色命名的跟踪器（CN）

现在最先进的跟踪器或是使用光照强度（RGB 值）或是使用纹理信息。尽管现在在视觉跟踪方面已经取得了很大的进展，但是对于颜色信息的使用还是仅限于简单的颜色空间转换。和视觉跟踪不同的是，在目标检测方面，复杂的，巧妙设计的颜色特征显示了非常好的效果，而利用颜色信息做视觉跟踪是一件很难的事情。颜色测量结果在整个图片序列中变化很大，原因包括光照改变，阴影，相机和目标几何位置的变化。此种方法建立在 CSK[28]跟踪器之上，CSK 是在一个单独的图像碎片中从目标中得到核心的最小方形分类器[29]。大体上，大部分运动跟踪都是通过查找两个相邻帧的相互关系，再确定目标对象的运动方向，无限迭代后，能完整地跟踪对象，CSK 也是如此。在确定跟踪对象后，根据目标位置扣出该帧的目标窗和下一帧的目标窗，再对这两个窗进行 FFT，转化后在频域图直接点乘。这个过程可简单理解为是求两个相连帧的频域共振位置，然后将共振频域图利用核函数进行核映射，再进行训练。训练过程引入原始响应 Y，Y 可以理解为是对象的起始位置,因为起始位置都是第一帧的中心，所以能看到 Y 的图像是根据跟踪窗大小的建立的高斯函数。训练的目的则是要找出当前帧对应的α，其训练公式如式（3-3）：

$$\alpha = \mathrm{F}^{-1}\left(\frac{F(y)}{F(k)+\lambda}\right) \tag{3-3}$$

训练完毕后，根据下一帧的核映射，可检测出对应的响应图像y',检测公式为式（3-4）所示：

$$y' = F^{-1}(F(k) \odot F(\alpha)) \tag{3-4}$$

CSK 算法的速度很快，其原因是其使用了循环结构进行相邻帧的相关性检测的。所谓的循环结构，其实质是两帧在频域上进行点乘操作，即两帧在时域上做卷积。在以前的运动跟踪，相关性检测都是使用滑窗法完成的，若窗的滑动步长为 1，即可看作是两帧之间做卷积。但在时域上做卷积的运算量是非常惊人的，而在频域上做点乘，运算量则小得多，因而 CSK 使用的循环结构在频域上做分析，则能很好地提速。

自适应的颜色算法（CN）先将 RGB 空间的图像映射到 CN 空间（CN 空间是

11 通道，分别是 black, blue, brown, grey, green, orange, pink, purple, red, white, yellow），并对每一个通道均进行 FFT、核映射，最后将 11 通道的频域信号线性相加，继而完成 CSK 的 计算，如α的计算、训练、检测等。

3.3.3 基于梯度检测和颜色跟踪的检测跟踪算法

由 3.3.1 和 3.3.2 已经介绍了一种目标识别的算法和一种跟踪的算法，由于自动视频监控需要进行较长时间的跟踪，并且需要有很强的鲁棒性，然而监控视频的效果却并不是那么理想，经常有跳帧的现象，基于此，于是希望通过融合这两种优秀的目标识别和跟踪算法来实现一个新的检测跟踪算法，其流程图如下：

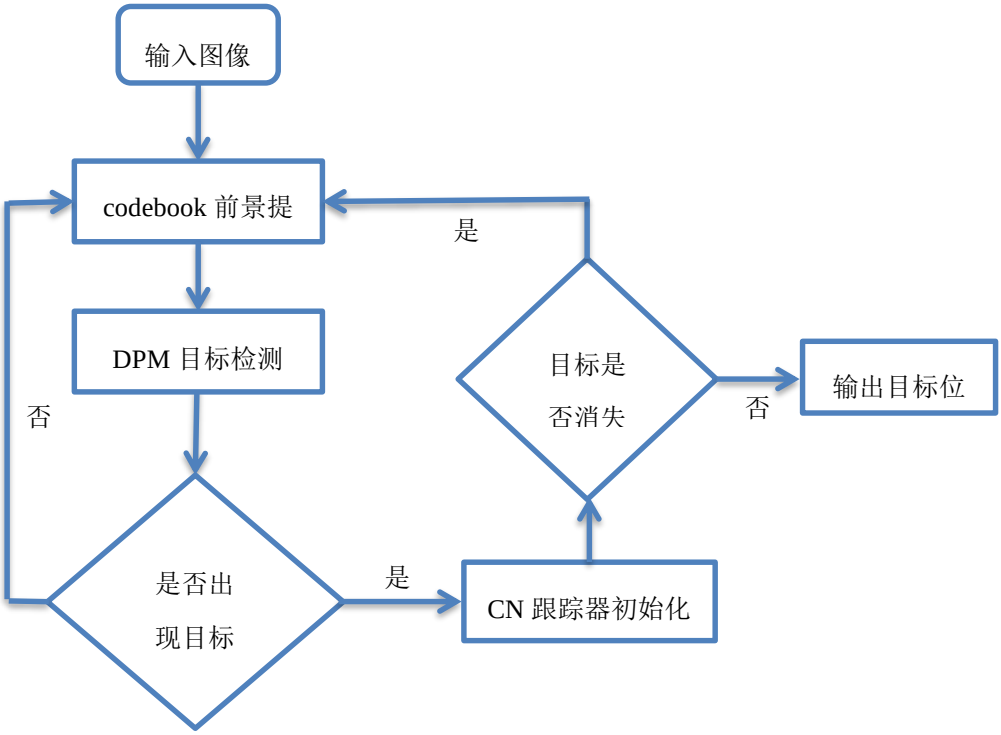

图 3.4 基于梯度检测和颜色跟踪的检测跟踪算法流程图

由于跟踪时候目标会随着离镜头的远近而大小的变化，所以在检测的时候还会定时检测用来更新目标框的大小。而且由于考场一般不仅仅只有一个老师，需要进行多目标的跟踪，在数据关联上我们使用一些几何上的技巧来进行判定。通过新的目标框与上一次的目标框的相关性来进行调整。其公式如式（3-5）所示：

$$Score[i] = I[i] + C[i] + A[i] \tag{3-5}$$

其中 *Score*[i]是相似度得分，*I*[i]为相交面积的比例，*C*[i]和 *A*[i]为二个矩形框的中心距离以及中心夹角的度数，每次在数据关联的时候都进行一次相似度得分的计

算来确定数据的关联。

3.4 实验结果分析

为了验证算法的效果，本次实验是对真实的考场数据进行跟踪，其中算法的实现是用 C++进行编写，视频的分辨率为 720*480。其主要背景为课桌的背景，视频总长为 2600 帧。首先通过背景减除算法将前景提取出来，其效果如下：

图 3.5 第一张为背景减除结果，第二张为经过形态学处理后的结果，第三张为在前景区域检测目标的结果。

图 3.6 跟踪结果图，从左到右的图像依次为 80、500、1070、1640、1700、1730 帧。蓝色方框为跟踪框。

由于摄像头比较老，视频效果非常不好，所以在单纯的使用 CN 进行跟踪的时候会经常目标丢失，而且还有一个问题就是此摄像头在传输的时候会掉帧，这样导致的结果就是人物会进行瞬间的移动（可见图中的第 4、5 张图片），在这种情况

下通过不断的检测和匹配，可以看到在第 6 张图片中目标又跟踪回来，并且可以看到，目标框是会根据目标在视频中的大小变化而变化的，这也是这个算法的好处之一。

由于用到了 DPM 算法，DPM 的速度还有待提高，此跟踪算法的 FPS 为 25，勉强能达到实时的要求。

3.5 本章小结

本章首先介绍了目标跟踪算法的应用方向，然后讲述了目标跟踪算法的分类，主要介绍了判别类的跟踪算法，然后详细讲述了基于检测的跟踪算法，最后提出了一种新的基于检测的跟踪算法，该算法结合了 2 种优秀的检测和跟踪算法，可以进行较长时间的鲁棒的跟踪。

经过试验证明，本章提出的基于梯度的颜色检测跟踪算法能够较好的跟踪目标，并且能够在条件恶劣、分辨率较低的情况下持续的跟踪目标，即使是出现了跳帧等极端情况，该算法也能够迅速的重新找到目标并继续跟踪，并且该算法还能够根据目标的大小变化而改变目标框的大小，由于是基于检测的跟踪算法，该算法在出现遮挡，消失一段时间之后也能够继续的跟踪目标，所以有较强的鲁棒性。在效率上，该算法能够达到 25FPS，能够满足实时性的要求，不过只是勉强，后续的工作可以在提升算法速度上加以研究。

第 4 章 基于混合特征的动作识别算法研究

4.1 引言

人作为社会的主体，永远是研究的重点，在计算机视觉领域，人的动作识别已经被研究多年，其应用场景非常广，现在摄像头如此之多，比如对监控视频中的人物进行动作的识别判定是否有违规行为，或者对网上的在线视频进行分析，
对电影中的人物动作进行分析，并且由于最近 VR、AR 等非常火热，各种各样的体验产品层出不穷，由于其需要沉浸入一个虚拟的世界，所以传统的交互方式已经不适合，而其中动作识别(如手势识别)将成为其交互的重要方式，这种方式是非常自然的。由于对人的关注的意义，动作识别对计算机或者机器理解人类社会有着非同凡响的重要性，无论是对高层语义的分析还是对整个场景的描述，人在其中都是最重要的研究对象[30]。

动作识别作为计算机视觉领域的高层应用，其包含了很多方向，如图像处理、信号处理、模式识别等相关技术，目前主要的动作识别算法分为二大类，第一类是基于局部特征的识别方法，比如 HOG3D[14]、HOF[15]、DT[16]，还有基于姿态的行为识别方法，姿态可以提取更加细节的信息，如轮廓之类的[31]，然后通过关键姿势来区分不同的动作从而识别。

4.2 学生前景检测

在识别动作之前需要对动作的目标进行定位，在这里我们使用一种深度神经网络的方法 Faster-Rcnn[32]来对学生进行定位。Faster-Rcnn 是比 Fast-Rcnn[33]更快的一种 Regin Cnn[34]。方法，其主要利用了 CNN 在图像识别领域获得的巨大成功，于是希望能够对图像目标在进行识别的同时也能够进行定位，这样就产生了 Regin Cnn，这种方法不仅能够准确的识别出图片中的物体是什么，并且能够定位目标物体的位置，这样就可以做目标的检测，在目前所有的目标检测算法中，利用神经网络来做的方法要远远好于传统的利用人工特征的信息来检测的算法。

该算法能够准确的定位学生的位置，并且给出相关位置的方框，这样有效的帮助了后面进行动作检测时所需要的目标位置，该算法的优势在于其定位非常的准确，但其缺点也比较明显，就是需要有训练数据，并且在检测的时候速度也有一定的消耗。

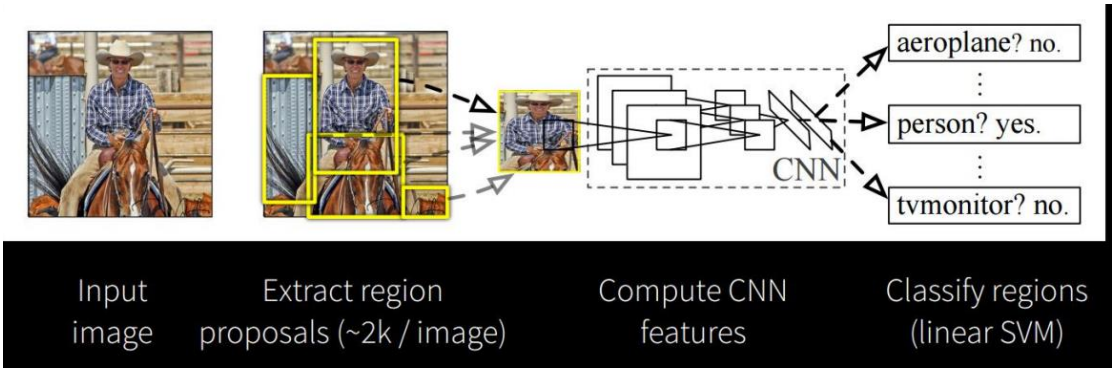

图 4.1 Rcnn 框架[34]

其步骤为：

(1)输入测试图像；

(2)利用区域选择算法在图像中从上到下提取 2000 个左右的候选区域；

(3)将每个候选区域缩放（warp）成 227*227 的大小并输入到 CNN，将 CNN 的 fc7 层的输出作为特征；

(4)将每个候选区域提取的 CNN 特征输入到 SVM 进行分类；

(5)对于 SVM 分好类的候选区域做边框回归，用 Bounding box 回归值校正原来的建议窗口，生成预测窗口坐标。

Rcnn 的目标定位准确率很高，但是由于需要产生 2000 多个候选框，并且都要进行 CNN 的提取，这样的时间消耗是非常大的，于是 Fast-Rcnn 和 Faster-Rcnn 被提了出来，Fast-Rcnn 改进的地方是将整张图片进行 CNN 的训练，最后再用候选框来局部进行特征的提取，这样就避免了对 2000 多个候选框区域做 CNN 操作，时间减少了不少，然而还是需要在 CNN 卷积结束之后再对 2000 个框分开进行提取特征，其中 2000 个框的得到也需要花费时间。Faster-Rcnn 的主要思想还是和 Rcnn 相同，但是在速度上改进了许多，其候选框的提取通过 RPN 来进行，并且产生的候选框的 CNN 和目标检测的 CNN 共享，这样就并行了候选框提取和 CNN 计算。

在本数据集上测试 Faster-Rcnn 的速度为 2ms 一张。

4.3 混合特征研究

在基于局部特征的动作识别方法中，最重要的就是特征的设计与选取，在这里我们介绍几种常用的特征。

4.3.1 光流梯度特征

光流是一种非常重要的概念，其已经在动作和前景提取方面取得了很好的效

果，首先光流的计算公式如式（4-1）所示：

$$\mathrm{I}(\mathrm{x}+\Delta\mathrm{x},\mathrm{y}+\Delta\mathrm{y},\mathrm{t}+\Delta\mathrm{t})=\mathrm{I}(\mathrm{x},\mathrm{y},\mathrm{t})+\frac{\partial I}{\partial x}\Delta\mathrm{x}+\ \frac{\partial I}{\partial y}\Delta\mathrm{y}+\frac{\partial I}{\partial t}\Delta\mathrm{t} \tag{4-1}$$

其中(x,y,t)表示指定坐标在 t 时刻的瞬时速度，$\mathrm{I}(\mathrm{x},\mathrm{y},\mathrm{t})$为(x,y)在 t 时刻的光照强度。同时考虑到两帧相邻图像的位移足够短，则：

$$\mathrm{I}(\mathrm{x},\mathrm{y},\mathrm{t})=\mathrm{I}(\mathrm{x}+\Delta\mathrm{x},\mathrm{y}+\Delta\mathrm{y},\mathrm{t}+\Delta\mathrm{t}) \tag{4-2}$$

因此：

$$\frac{\partial I}{\partial x}\Delta\mathrm{x}+\ \frac{\partial I}{\partial y}\Delta\mathrm{y}+\frac{\partial I}{\partial t}\Delta\mathrm{t}=0 \tag{4-3}$$

$$\frac{\partial I}{\partial x}\frac{\Delta\mathrm{x}}{\Delta\mathrm{t}}+\ \frac{\partial I}{\partial y}\frac{\Delta\mathrm{y}}{\Delta\mathrm{t}}+\frac{\partial I}{\partial t}\frac{\Delta\mathrm{t}}{\Delta\mathrm{t}}=0 \tag{4-4}$$

最终可以得到

$$\frac{\partial I}{\partial x}V_x+\ \frac{\partial I}{\partial y}V_y+\frac{\partial I}{\partial t}=0$$

这里的V_x、V_y是 x 和 y 方向的速率，即称为$\mathrm{I}(\mathrm{x},\mathrm{y},\mathrm{t})$的光流，$\frac{\partial I}{\partial x}$、$\frac{\partial I}{\partial y}$、$\frac{\partial I}{\partial t}$是图像(x,y,t)在 t 时刻特定方向的偏导数。I_x、I_y、I_t的关系可以用下式表示：

$$I_xV_x+I_yV_y=-I_t \tag{4-5}$$

在计算得到每一帧的光流场(x,y)之后，将图像进行区域的划分，并且对每个区域计算统计直方图。一般将每一帧分为 9 个区域，每个区域将 360 度分为 9 个区域，也就是每个扇形区域有 40 度，对于一片视频，一般分为 4 段，如下图所示：

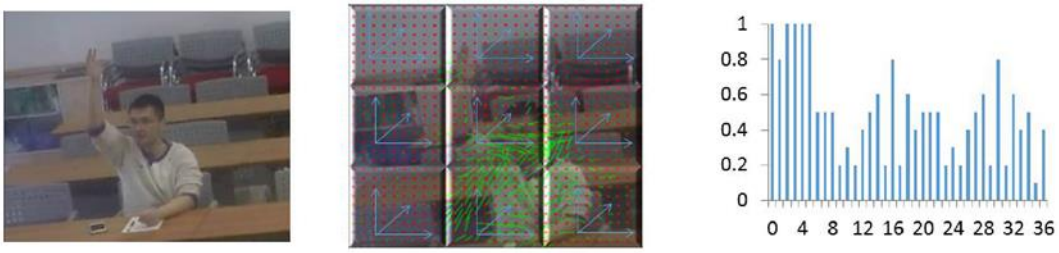

图 4.2 HOF 特征提取示意图,绿色箭头为光流场

这样计算下来一个视频所有的特征向量长度为 9*9*4=254 维，HOF 特征的好处在于其利用的是光流信息。此信息记录了动作在时间上的变化，比如向左伸手和

向右伸手，其在时间上的移动方向是不一样的，通过 HOF 特征就能够很好的区分这类动作.但是由于其仅仅只考虑了时间上的移动，而缺乏空间和纹理信息，所以并不能很好的区分左举手和右举手等动作。

4.3.2 梯度直方图特征

和 4.3.1 节所提到的 HOF 特征不同，这里的梯度直方图特征(HOG)计算的是图像的梯度，对于图像中像素点(x，y)：

$$G_x(x,y) = H(x+1,y) - H(x-1,y)$$

$$G_y(x,y) = H(x,y+1) - H(x,y-1) \tag{4-6}$$

式中$G_x(x,y)$、$G_y(x,y)$、H(x,y)分别表示输入图像中的像素点(x，y)的 x 方向梯度、y 方向梯度和像素值，这样可以得到该点的梯度的大小和方向如式（4-7）和（4-8）所示：

$$\mathrm{G(x,y)} = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{4-7}$$

$$\alpha(\mathrm{x,y}) = \tan^{-1}\frac{G_y(x,y)}{G_x(x,y)} \tag{4-8}$$

在得到梯度信息之后即可以进行直方图的计算，其方式和 4.3.1 节中的直方图的计算一样，由于梯度信息记录的是图像的纹理变化，即空间位置的区别，所以此特征能够很好的区分由于空间位置的区别而区别的动作，例如举左手和举右手等，但是缺少了时间上的移动信息。

4.3.3 移动区域直方图特征

在 4.3.1 和 4.3.2 节中提到的特征都有各自的有点，但也有各自的缺点，要么是缺少了空间的信息，要么是缺少了时间上的信息，而移动区域直方图特征(MBH)很好的综合了二者的优点，其通过对视频图像提取光流场的方式将动作的时间信息考虑进去，然后再对提取的光流场图像进行一次梯度直方图的提取，这样做就又将动作的空间轮廓信息考虑进去，使得该算法能够将动作从时间和空间上分别区分开来，达到了很好的区分率。

其特征的设计为，先将图像帧计算出光流场的信息，这样可以得到一张 x 方向的光流场图像和一张 y 方向的光流场图像，然后接着分别在这 2 张图像上提取 HOG 特征，最后进行直方图的统计，由于统计了空间上和时间上的信息，所以该特征能够比较好的区分动作，并且在识别的时候能够比较快速的进行计算，特征维度不是

很高，其效果如图 4.3 所示：

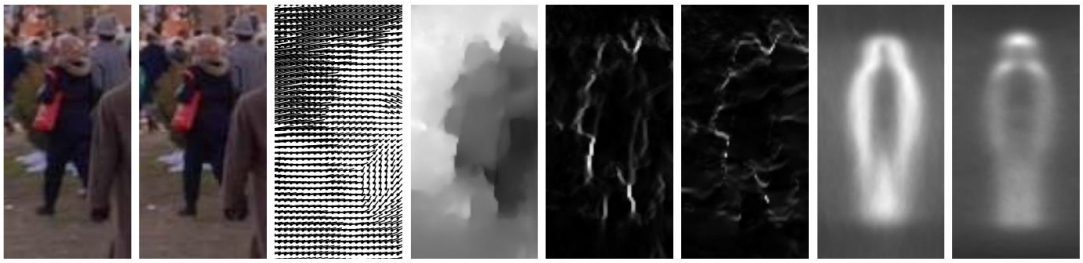

图 4.3 MBH 特征示意图[26]

其中 a 和 b 为视频的 t 和 t+1 帧，c 为光流场的方向示意图，d 为光流场大小的示意图，e 和 f 分别为在 x 和 y 方向上对光流场提取 HOG 特征的图像，g 和 h 为整个在训练集训练后得到的目标 MBH 特征可以看到 MBH 特征很好的考虑到了时间上的变化也考虑到了空间上的变化，所以其对动作的区分度效果很好。

4.4 支撑向量机

在得到特征之后需要将特征进行训练，本文选择支撑向量机（SVM））作为分类器，支持向量机（SVM）最早诞生在 90 年代，其是属于统计学习的范畴，基本思想是通过优化最小风险函数来对目标进行分类，其特点是对统计样本并不是很高，即使是统计样本较少，也可以得到比较好的结果。

这里支撑向量机的基本推导就不再赘述，主要讲一下本文支撑向量机选择的核函数为线性核函数，其如式（4-9）所示：

$$\mathrm{K}(x_i, x_j) = x_i{}^T x_j \tag{4-9}$$

本文并没有选择一些很复杂的分类器因为我们的数据维度很高，一般都有几百维，而所有的数据集合才 3000 多个(数据会在第五章详细介绍)，所以我们认为其应该很有可能在线性核函数下的效果会好一些，实验结果也表明确实如此，在线性核函数下我们的准确率能达到 94.16%，而在多项式核函数下仅有 90.22%，在 RBF 核函数下仅有 92%，比线性核函数都要差一些。

4.5 实验结果分析

本文的实验都是在自己拍摄的数据集上进行的，其具体情况可以在后面一章

中会有详细介绍，实验的机器与上文一样保持不变。

4.5.1 学生位置检测结果分析

这里使用的是 Faster-Rcnn 框架来进行的识别，首先进行训练，我们选取了 200 张坐着的学生作为正样本，而负样本为随机截取的不包括学生的图片。最后的结果如图 4.4 和图 4.5：

图 4.4　Faster-Rcnn 检测学生结果展示

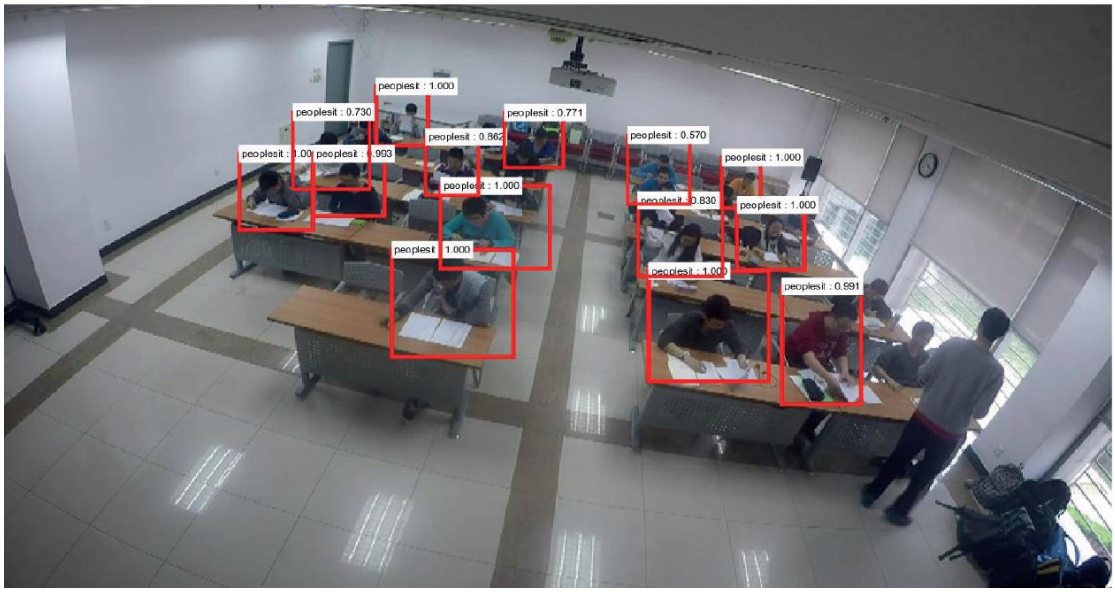

图 4.5　Faster-Rcnn 检测结果展示

可以看到检测的结果还是可以的，并且相似度也挺高，虽然有部分同学因为太小或者被遮挡而不能被检测出来，但是总体来看这个结果是可以接受的，一共有 24 个同学，而算法框出了其中的 14 个，并且这个算法的时间消耗很小，可以达到 2ms 每张，可以满足实时性的要求，这里的结果将直接作为后续动作识别的输入之一，为后面的动作识别奠定了基础。

4.5.2 动作识别结果分析

这里的数据主要来自于本文所建立的 1300 个视频剪辑，一共有 9 种动作，每一种动作都至少有 110 个视频小片段，其结果展示如下图所示：

表 4.1 各个特征检测准确率

	HOG	HOF	MBHx	MBHy	MBH	iDT	C3D
N	54.54	63.10	83.42	86.09	91.97	53.14	33.25
PL	47.27	98.18	100	98.18	100	63.63	45.88
PR	38.59	98.24	97.36	97.36	100	61.40	55.21
RL	23.33	94.16	95.83	90	95.83	43.33	35.22
RR	33.83	91.72	95.48	96.24	97.74	28.03	50.15
S	44.11	97.79	98.52	98.52	99.26	58.82	35.16
SL	25	95.13	94.44	93.05	95.83	52.08	44.28
SR	31.93	88.23	88.23	92.43	92.43	53.78	29.25
TL	16.66	81.57	80.70	56.14	85.08	25	47.21
TR	13.22	54.54	72.72	52.89	83.47	9.3	27.69
AVG	33.98	85.28	90.44	86.44	94.14	44.68	40.33

在这里我们一共有九种动作，其中 N：正常动作 PL(PR)：向左(右捡东西)RL(RR)：向左(右)举手 S：起立 SL(SR)：向左(右)伸手 TL(TR)：向左(右)转头。

该结果使用的是交叉验证方式，一共将 1300 个视频分为 5 组，即每个特征进行 5 此测试，每一次测试都有 1040 个训练数据和 260 个测试数据。

可以看到 MBH 的结果最好，可以达到 94.14%，并且在特别疑似作弊行为的动作如站立以及向下捡东西和伸手等识别率都非常高，然而可以看到 C3D 这种深度学习的特征效果却不好，很大的原因可能是训练数据太少，其并没有学习到很好的特征，其他的如 HOG 和 HOF 等特征相对而言就中规中矩。

4.6 本章总结

本章在开始介绍了动作识别的研究意义，然后着重介绍了基于局部特征判别的算法框架，接着介绍了几种比较好的特征，其中 HOG 和 HOF 特征各有优缺点，一个在空间上有不俗的表现，一个在时间上有好的表现，最后将二者结合使用了一种 MBH 特征来作为分类器分类的依据，经过试验证明该特征是一个非常好的特征，其能够很好的区别 9 种测试动作，特别是那种非常疑似作弊动作的行为如向

下弯腰和向左右伸手等等，识别率超过了 94%。

在本章中还介绍了一种基于深度神经网络的目标检测算法，该算法主要目的是为了能够自动的提取坐着的学生，为以后的动作识别作为铺垫，该算法准确率非常高，并且速度也很快，达到了 2ms 一帧，效果显著。

在本文中有提到一个基于深度学习的动作特征提取方法，然而此方法在实验中表现一般，但鉴于深度学习在目标检测和图片识别方向有如此大的发展，今后的动作识别方向也应该是往这方面走，通过深度神经网络来学习的特征最后肯定能比人工设计的特征效果要好。但是目前而言还没有比较好的针对视频流的神经网络架构，如果能够设计出比较好的神经网络架构来对视频流进行优化的话，应该是可以做出比较好的结果的。

第 5 章 考场自动监控系统框架

5.1 考场自动监控系统框架

考场一直是国家教育的重要检验场，考试肩负了国家选拔人才，社会公平的职责，其意义重大，然而近几年来有一些学生、老师和家长为了一己私利在考场内作弊，甚至是团伙作弊，监考老师和学生串通起来一起作弊，这种行为严重影响了社会的公平性，造成了恶劣的影响。然而随着科技的发展，计算机视觉的不断强大，这种场景下实现自动监控成为了一种可能，通过对老师的跟踪以及学生动作的识别，对一些可疑的动作进行报警，可以作为一种防止作弊，增加公平性的措施，让那些想不劳而获的人心生畏惧，达到考试的公平性。

本文提出的考场自动监控系统框架主要包含 2 个部分，一个是老师的跟踪，另外一个是学生动作的识别，在前面二、三、四章中已经详细描述了相关算法，在这里就不再赘述，图 5.1 是考场自动监控系统框架图：

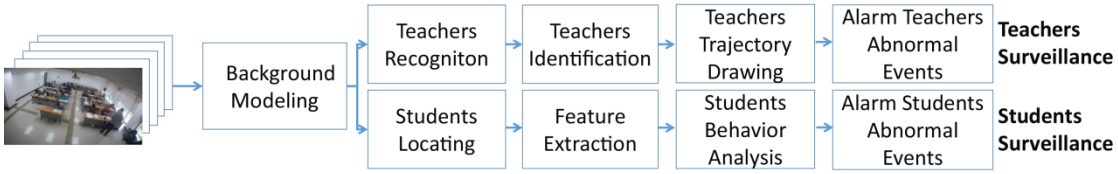

图 5.1 考场自动监控系统框架图

对于老师轨迹的跟踪，由于考试的时候一般都是有规定老师在特定时间的位置，所以可以根据相关规定，对老师的位置进行判别，如果其在某个时候不在要求的位置，那么可以进行报警提醒。并且在跟踪的时候需要对老师与学生的位置进行判定，因为有可能学生和老师一起作弊，所以需要对老师是否进入学生答题区域进行一个判定。

在学生方面，学生的动作识别可以对相关学生起到警示的作用，并且如果出现疑似作弊行为，则在视频中提醒相关监控人员及时注意，这样可以更好的对考试公平性做出保证。

由于二者是考场中仅有的目标，在这里可以进行并行的设计，对于考生可以单独占有一台或几台 CPU，对于老师跟踪也是如此，它们之间是相互独立的，这样有助于对系统速度的提升，最后进行报警的时候统一进行，可以放在一个窗口中同时展示。

图 5.2 为本系统学生异常动作报警示意图：

图 5.2 学生异常动作报警示意图

可以看到那些框坐着的学生的框是通过 Faster-Rcnn 得到的，然后对框进行动作识别的处理，当识别出有异常动作的时候，其会将检测有异常的框放大后展示在屏幕左下角，并且在右边配上异常动作的名称，这样可以方便监视人员进行合理的判别。

5.2 考场数据集

本文还有一个重要的贡献就是拍摄了大约 6 个小时的考场视频，我们手里有真实的考场视频，然而在观察中发现里面的异常动作非常少，在如今大部分好算法都需要大量数据来进行学习的情况下，这个问题显的很棘手，于是我们进行了数据的拍摄与标注。在动作识别领域已经有一些数据集，比较有名的有 KTH[35]、Weizmann[36]、IXMAS[37]、Hollywood[38]、UCF[39]、The Olympic sports[40]、HMDB51[41]等。

KTH 数据集：数据集于 2004 年的发布,是计算机视觉领域的一个里程碑。数据库包括有 4 个不同场景，参与的人数有 25 个人，总共的动作有 6 类动作，一共有共计 2391 个视频样本，其视频样本中包含了衣着的变化、尺度的变化以及一些光照的变化，摄像机是单一摄像机，背景为纯白色背景。

Weizmann 数据库：数据集包含 10 个动作（run, bend, jack, jump, walk, wave1,

wave2, side, skip, pjump），每种动作都有 9 个不同的样本。视频的摄像头为单一摄像头，并且背景简单，为纯色背景，该数据集还标注了一些前景和背景信息。

IXMAS 数据库：数据集有五个视角，它们分别为室内的四个方向以及顶上，其中场景比较简单，光照基本不变，背景为纯色。其中包含 14 个动作，每个动作都重复在一个视频中做 3 次，一共有 11 个人。

UCF 数据库：数据集样本主要来自网络 YouTube 等视频网站，视频的源文件主要来自广播电台，此数据集主要是运动样本，其中 UCF101 是使用的最多也是数据量最大的一个数据集，其样本为 13320 段视频，有 101 类。

Olympic Sports 数据集：该数据集也是运动类别的数据库，其主要视频来源为视频网站 YouTube，其中一共有 50 度个视频，一共有 16 个运动类别，每一部视频都标记了运动的类别。

HMDB51 数据集：基本上都是截取的电影，其中很大一部分来自于视频网站 YouTube，该数据库一共有 6849 段样本,有 51 类，每类都至少包含有 101 段样本。

可以看到上面的数据集都包含有很多动作，我们数据集一共包含有 9 个动作，它们分别是{Pick up left、Pick up right、Raise left hand、Raise right hand、Stand up、Stretch hand left、Stretch hand right、Turn around left、Ture around right}。其具体数目如下表：

表 5.1 考场数据集动作信息

	PL	PR	RL	RR	S	SL	SR	TL	TR	N
数量	110	114	120	133	133	144	119	114	121	187

其中每段视频的帧数大概在 30 帧-60 帧之间，拍摄人物一共有 24 个，拍摄角度一共有四个角度。总共标记的视频有 1300 个，每一个基本上都有清晰的人物轮廓，每一个类别的视频都至少包含有 110 个样本，每一个样本都含有一个学生做一个完整的动作。其基本视角如图 5.3 所示：

图 5.3 考场监控视频数据视角图

其中具体的标注的视频图 5.4 所示，分别选取了 9 类动作中的一个样本进行动作切分的截图：

图 5.4 考场监控数据集动作集合

数据集标注的情况都是比较好的，很少有重叠和覆盖的现象，此数据集可以

用来进行考场动作视频识别的相关训练和检测工作，由于动作种类较多，并且人员也足够多，每一类动作的数量也还可以，对比与上面所提到的一些已经存在的数据集，总的而言可以满足一般研究的需求。

5.3 本章总结

本章主要介绍了基于前面二、三、四章所提到的技术所提出的一个考场监控视频系统的框架，并且针对考场视频中异常动作很少的问题，建立了自己的数据集，该数据集拥有 4 个角度，总长度达到 4 个小时，并且标注了 1300 个动作视频共 9 类动作，每一类动作都至少有 110 个视频数据，该数据集对于考场视频监控这一特定场景的研究有一定的意义。

第 6 章 总结与展望

6.1 论文工作总结

本文的主要研究了监控视频中的前背景提取、多目标跟踪以及人体动作的识别算法技术。对于这些技术从前人的工作中进行学习、比较，循序渐进的展现了一些现有算法的优点以及不足，然后根据实际考场的应用情况，对这些算法提出一些改进。

首先在前背景检测方面介绍了现在主流的前背景分割算法，比较了他们之间的优劣，然后针对考场视频监控这样一个特定场景，选择了一种基于码本的前背景提取算法，这种算法有效的解决了动态背景的问题，能够针对监控摄像头这种老旧的设备，得到我们想要的运动目标，能够有效的避免高光和阴影的影响，并且效率也非常高，能够满足监控视频实时的要求。此外，在动作识别的时候还对基于卷积神经网络的目标定位进行了研究，其有效的解决了在动作识别中自动找到目标物体的问题，并且达到了实时的要求。

在多目标跟踪方面主要介绍了基于检测的的跟踪算法，并且受到这种算法的启发，针对需要对考场的老师进行长时间的跟踪，提出了一种混合 DPM 和 CN 的检测跟踪算法，该算法能够有效的解决因为摄像头老化的问题而出现的跳帧现象，在即使发生跳帧（即目标物体会瞬间移动一段唯一）现象时，也能够及时的跟踪回到目标物体。并且还通过几何学的技巧将单目标跟踪扩展到了多目标跟踪。

在动作识别方面，主要介绍了基于局部特征的动作识别方法，并且介绍了几种常见有效的特征 HOG、HOF、MBH 等以及基于深度神经网络的 C3D 特征，然后通过实验比较得出了 MBH 在本数据集中的效果最好，而 HOG 和 HOF 特征由于只考虑了时间或者空间的一个部分，所以得到的效果略差，而深度学习的特征 C3D 的表现却不尽如人意，目前在视频方面，深度学习还并没有展现其威力。

本文还针对考场这一特定的场景建立了自己的数据集，该数据集一共有 4 个角度共 6 个小时的视频，并且针对考场的动作，本文进行了标注，一共标注了 9 类动作，每一类动作都包含至少 110 个短的视频段，该数据集很好的填补了在动作识别中考场这一特殊场景的空白，由于真实考场场景中异常动作十分少见，这个数据集可以用来进行一些基于机器学习和深度学习的动作识别的探索。

本文所提出的针对考场监控视频的系统框架能够对考场中的 2 类重要目标学生和老师进行自动处理，有一定的自动监控能力。

6.2 未来工作的展望

本文对监控视频的自动分析系统的具体技术进行了一些研究，并且提出了一个针对考场的监控视频的自动分析系统，主要针对老师的轨迹以及学生的动作进行了一个判别。然而在真实的场景中很有可能老师和学生一起作弊，他们之间可能会有交互的动作，目前所提出的系统只能够通过越界检测来判别老师和学生是否进行了交互，而并没有当做一个整体来考虑。在未来可以在动作识别方面研究多人物互动动作的研究，这样可以使得在判别老师和学生进行交互的时候有一个更准确的判断。

此外，在目标跟踪方面，目前的算法的速度还有待提高，由于是直接将 DPM 和 CN 二者进行的一个串联操作，虽然准确率很高，并且容错度也很好，但是时间消耗上有一点大，因为 DPM 每一次检测都需要一定的时间，在未来的研究方向可以考虑一下优化它们二者的速度，使用一些并行的技巧之类的。

本文提出的监控视频的自动分析主要针对了考场这样一个特定的场景，然而目前在计算机视觉领域中还没有一个很好的泛化的自动分析系统，主要是没有一个非常鲁棒能够在任何场景都能够跟踪好或者识别好的算法，未来的研究方向可以研究一下能够在更多场景应用的跟踪或者识别算法，到那个时候，计算机视觉才真正实现了其初始的目标：让机器拥有眼睛能够“看见”这个世界。

参考文献

[1] Ridder C, Munkelt O, Kirchner H. Adaptive Background Estimation and Foreground Detection Using Kalman-Filtering[C] Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on. IEEE, 1998:390-395.

[2] Davis J W, Bobick A F. The Representation and Recognition of Action Using Temporal Templates[J]. Proc of Cvpr, 2000, 23(3):928--934.

[3] Elgammal A, Harwood D, Davis L. Non-parametric Model for Background Subtraction[J]. Lecture Notes in Computer Science, 2000, 1843:751-767.

[4] Stauffer C, Grimson WEL. Adaptive Background Mixture Models for Real-time Tracking. IEEE International Conference on Computer Vision and Pattern Recognition 1999;2:246–52.

[5] Lee DS, Hull JJ, Erol B. A Bayesian Framework for Gaussian Mixture Background Modeling. IEEE International Conference on Image Processing 2003

[6] Harville M.A Framework for High-level Feedback to Adaptive, Perpixel, Mixture-of-gaussian Background Models. European Conference on Computer Vision 2002;3:543–60.

[7] Javed O, Shafique K, Shah M. A Hierarchical Approach to Robust Background Subtraction Using Color and Gradient Information. IEEE Workshop on Motion and Video Computing (MOTION'02); 2002.

[8] Kass, M., Witkin, A., and Terz, D. 1988. Snakes: Active Contour Models. Int. J. Comput. Vision 1, 321–332.

[9] Broida T J, Chellappa R. Kinematics and Structure of A Rigid Object from A Sequence of Noisy Images[J]. Proc Workshop on Marion Represenrarion & Analysis, 1986, 13(6):497 - 513.

[10] Comaniciu D, Meer P. Mean Shift: A Robust Approach Toward Feature Space Analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(5):603-619.

[11] Hare S, Saffari A, Torr P H S. Struck: Structured Output Tracking With Kernels[C] IEEE International Conference on Computer Vision. IEEE, 2012:263-270.

[12] Kalal Z, Mikolajczyk K, Matas J. Tracking-Learning-Detection[M]. IEEE Computer Society, 2012.

[13] Dollar P, Rabaud V, Cottrell G, et al. Behavior Recognition Via Sparse Spatio-temporal Features[C] Joint IEEE International Workshop on Visual Surveillance and PERFORMANCE Evaluation of Tracking and Surveillance. IEEE, 2006:65-72.

[14] Kläser A, Marszalek M, Schmid C. A Spatio-Temporal Descriptor Based on 3D-Gradients[C] British Machine Vision Conference 2008, Leeds, September. DBLP, 2008.

[15] Laptev I, Marszalek M, Schmid C, et al. Learning Realistic Human Actions From Movies[C] Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008:1-8.

[16] Wang H, Kl äser A, Schmid C. Dense Trajectories and Motion Boundary Descriptors for Action Recognition[J]. International Journal of Computer Vision, 2013, 103(1):60-79.

[17] Taylor G W, Fergus R, Lecun Y, et al. Convolutional Learning of Spatio-temporal Features[C] Proc. European Conference on Computer Vision. 2010:140-153.

[18] Du T, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[J]. 2015:4489-4497.

[19] Wren CR, Azarbayejani A, Darrell T, Pentland A. Pfinder: Realtime Tracking of The Human Body. IEEE Transactions on Pattern Analysis and Machine Intelligence 1997;19(7):780- 5.

[20] Horprasert T, Harwood D, Davis LS. A Statistical Approach for Real-time Robust Background Subtraction and Shadow Detection. IEEE Frame-Rate Applications Workshop, Kerkyra, Greece; 1999.

[21] Stauffer C, Grimson WEL. Adaptive Background Mixture Models for Real-time Rracking. IEEE International Conference on Computer Vision and Pattern Recognition 1999;2:246–52.

[22] Greiffenhagen M, Ramesh V, Comaniciu D, Niemann H. Statistical Modeling and Performance Characterization of A Realtime Dual Camera Surveillance System. Proceedings of International Conference on Computer Vision and Pattern Recognition 2000;2:335–42.

[23] Kim, Kyungnam, Chalidabhongse, Thanarat H, Harwood, David. Real-time foreground-background segmentation using codebook model[J]. Real-Time Imaging, 2005, 11(3):172-185.

[24] Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object Detection with Discriminatively Trained Part-Based Models[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(9):1627.

[25] Danelljan M, Khan F S, Felsberg M, et al. Adaptive Color Attributes for Real-Time Visual Tracking[C] IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014:1090-1097.

[26] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C] Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893 vol. 1.

[27] Lowe D G. Object Recognition from Local Scale-Invariant Features[C] International Conference on Computer Vision. IEEE Computer Society, 1999:1150.

[28] Henriques J, o F, Caseiro R, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[M] Computer Vision – ECCV 2012. Springer Berlin Heidelberg, 2012:702-715.

[29] Luo W, Xing J, Zhang X, et al. Multiple Object Tracking: A Literature Review[J]. Eprint Arxiv, 2015.

[30] Hemayed E, Bebars A A. A Survey on Vision-based Human Action Recognition[C] Image and Vision Computing 2010[J]. 2014.

[31] Smedt Q D, Wannous H, Vandeborre J P. Skeleton-Based Dynamic Hand Gesture Recognition[C] IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, 2016:1206-1214.

[32] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, PP(99):1-1.

[33] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2014:580-587.

[34] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[35] Schuldt C, Laptev I, Caputo B. Recognizing Human Actions: A Local SVM Approach[C] International Conference on Pattern Recognition. IEEE, 2004:32-36 Vol.3.

[36] Blank M, Gorelick L, Shechtman E, et al. Actions as Space-time Shapes[C] Tenth IEEE International Conference on Computer Vision. IEEE Xplore, 2005:1395-1402 Vol. 2.

[37] Weinland D, Ronfard R, Boyer E. Free Viewpoint Action Recognition Using Motion History Volumes[J]. Computer Vision & Image Understanding, 2006, 104(2–3):249–257.

[38] M. Marszałek, I. Laptev, and C. Schmid, Actions in Context[C] in IEEE Conf. on Computer Vision & Pattern Recognition, pp. 2929 - 2936, June 2009.

[39] Soomro K, Zamir A R, Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. Computer Science, 2012.

[40] Niebles J C, Chen C, Feifei L, et al. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification[C]. European Conference on Computer Vision, 2010: 392-405.

[41] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A Large Video Database for Human Motion Recognition[C]. International Conference on Computer Vision, 2011: 2556-2563.

致谢

光阴似箭，随着论文的即将完成，我的研究生生涯也将要告一段落，回首这 3 年来的研究生涯，感觉有新奇、苦涩、感动与充实，在这三年我体会到了做学术的艰辛，看似简单的工作并不是随随便便都能完成的，我领略到了那种没有想法没有 idea 的苦闷，但也感受到了那苦尽甘来最终完成的欣慰。不经一番彻骨寒，怎得梅花扑鼻香，越是辛苦的背后，收获的也越发的甘甜。

首先我要感谢的是我的导师胡事民老师，他为我的论文严格把关，在开题和研究的过程中给予了我许多帮助，并且作为一名成功的学者，他态度严谨、做学术一丝不苟的态度深刻的影响了我，这也是我感受最深的一点，作为老师，他每天来的比谁都早，并且一有机会就到实验室来辅导我们的研究工作，对学生要求严格。我相信这种严谨勤奋的态度将在我以后的工作、学习中产生巨大动力，让我在以后的人生道路上，能够像胡老师一样，对待任何事情都能够以一颗严谨勤奋的心情去对待。同时，他还在生活上给予学生关怀，如果有困难，会给予学生尽可能多的帮助。

其次要感谢我的大师兄张方略，他在我的论文后续的修改以及研究中出现的一些细节问题都给予了很大的帮助，通过他在学习和研究上的经验方法，指导我学习的方法，让我能够在遇到问题时及时的找到解决方案，这种授人以渔的帮助让我受益匪浅。

还有一起研究工作的范若琛、邹定南、李秉俊、王铭轩等同学和我的室友们在课题研究中给予我的帮助。

还要感谢我的父母，是他们让我能够走到今天，可以自由的选择自己发展的道路和方向，我的一切都来自于我的父母，感谢他们。

最后感谢在百忙之中抽出时间审阅论文的各位专家教授，感谢参与我的硕士毕业论文答辩并给予我批评和指导。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，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，本学位论文的研究成果不包含任何他人享有著作权的内容。对本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。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1992 年 5 月 7 日出生于湖北省宜昌市。

2010 年 9 月考入华中科技大学计算机系，2014 年 7 月本科毕业并获得工学学士学位。

2014 年 9 月考入清华大学计算机系攻读计算机技术工程硕士至今。

发表的学术论文

[1] Zhang Q L, Zhang F L, Fan R C. Abnormal Action Detection and Recogniton for Examination Room. (CADDM, quarterly 在审中)