杭州电子科技大学

硕 士 学 位 论 文

题　　目 面向视频人脸检测的深度学习算法研究

研 究 生 陈雪婷

专　　业 信号与信息处理

指导教师 叶学义 副教授

完成日期 2016 年 3 月

杭州电子科技大学硕士学位论文

面向视频人脸检测的深度学习算法研究

研 究 生：陈雪婷

指导教师：叶学义 副教授

2016 年 3 月

Dissertation Submitted to Hangzhou Dianzi University

For the Degree of Master

Study on Deep Learning for Face Detection in Video

Candidate: Chen Xueting

Supervisor: Associate Prof. Ye Xueyi

March，2016

杭州电子科技大学
学位论文原创性声明和使用授权说明

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。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。申请学位论文与资料若有不实之处，本人承担一切相关责任。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

本人完全了解杭州电子科技大学关于保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属杭州电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为杭州电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密论文在解密后遵守此规定）

论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

摘 要

人脸检测是人脸识别、表情分析、人脸跟踪等人脸信息处理前提和基础。随着视频监控覆盖面的不断扩大，人脸监控所具有的不易被观测对象发现的显著优势使得视频人脸检测被越来越多的应用在了犯罪分析、智慧安防、人工智能等领域。现有的视频人脸检测算法在处理非理想条件（包括背景复杂、光照异常、人脸旋转等等）下的检测问题，往往仅针对其中某种情况有较好的效果，当多种非理想条件并存时，检测性能急速下降。而实际的视频信息中多种非理想条件并存是常态，因此针对这种复杂条件下的视频人脸检测，本文引入深度学习理论并结合视频帧间的连续性，探讨具有较强鲁棒性、误检率和漏检率低，且检测速度快的方法，以期为智能监控及智慧安防提供基础支持。具体研究内容如下：

首先，以深度学习理论和人脸检测神经网络为基础，提出一种级联型概率态受限玻尔兹曼机学习网络以实现视频单帧的人脸检测。它首先利用概率态受限玻尔兹曼机（Probability state-Restricted Boltzmann Machine，P-RBM）中神经元的概率表征来模拟人脑神经元所具有的连续分布激活状态，然后通过级联多个 P-RBM 构建深度学习检测网络来仿真人脑视觉系统对图像信息的层次学习模式，并以逐层递减隐藏层神经元数来控制网络规模，最后采用分层训练和整体优化的机制来缓解鲁棒性和准确性的矛盾。该算法利用学习网络在充分提取输入数据各层次特征的基础上建立从底层特征到高层语义的映射，继而获得输入数据的语义信息以准确地完成检测任务。

其次，上述这种面向视频单帧的人脸检测并未利用视频特有的帧间连续性信息。因此，在上述研究基础之上，进一步研究视频帧间的连续性，提出多帧间信息融合的视频人脸检测算法。针对视频单帧的人脸检测结果，它首先利用人脸肤色区域长宽比去除部分误检区域，其中长宽比允许范围的设定采用自适应更新方式以获得检测视频最适宜的边界条件，然后通过视频帧间人脸位置变化规律估计当前帧的检测结果，并与真实检测结果进行对比，利用对比规则对检测结果进行修正，删去误检区域，补上漏检区域，提高算法的检测准确率。

实验数据表明，面向视频单帧人脸检测的级联型 P-RBM 学习网络不仅能实现较低漏检率和误检率的检测效果，同时对旋转人脸的检测具有较强鲁棒性。另外，其检测速度较快，基本能满足实时检测的要求；将其与多帧间信息融合算法相结合实现的视频人脸检测不仅保持了原有的较快检测速度和较低误检率，同时显著降低了漏检率，还提高了对部分被遮挡人脸的检测性能。

关键字：视频人脸检测，深度学习，概率态受限玻尔兹曼机（P-RBM），多帧间信息融合

ABSTRACT

As the primary technology to extract information from face, face detection is the premise and foundation of face recognition, facial expression analysis and face tracking. At the same time, face monitoring has a significant advantage that it is not easy to be found by the monitored object. Thus, with the expansion of video surveillance coverage, video face detection has been more and more used in crime analysis, intelligent security, and artificial intelligence. However, many factors need to be considered in video face detection, such as the complex environment, the partially obscured face, the face rotation angle, and so on. In addition, the detection speed is also need to be taken into account because of the real-time requirement.

Point to these influence factors, a video face detection method based on the deep learning and the video's inter-frame continuity is proposed, which is trying to achieve stronger robustness against complicated backgrounds, illumination, rotation angles with the lower missing detection rate and the lower false detection rate, and provide theoretical support for intelligent monitoring and wisdom security. The concrete research content is as follows:

First of all, according to the theory of deep learning and face detection based on neural network, a cascaded learning network based on multi-layer Probability state-Restricted Boltzmann Machine(P-RBM) is proposed to realize face detection in a single frame video image. It first uses the probability state of the neurons in P-RBM as their activation degree, which better models the activity state's continuous distribution of the neurons in human brain. This design not only retains the weak active information, but also decreases the effect caused by the former layer's error. Secondly, this method simulates the hierarchical learning mode in human brain by cascading multiple P-RBMs. This cascaded network can realize the multi-layer nonlinear mapping and obtain the semantic feature of the input date. What's more, it can learn the relationship hiding within the data to make the learned features be more promotional and expressive. Simultaneously, the number of the hidden layer's neuron decreases layer-by-layer to control the network's scale and enhance the robustness. Finally, it uses the layered training and whole optimization to balance the robustness and accuracy.

This face detection in the single frame image method does not use the inter-frame continuity information, which is the unique advantage of the video. Thus, based on the above research, the continuity between video frames is to be further studied, and a video face detection method with multi-inter-frame information fusion is proposed. Firstly, the aspect ratio of the face skin color area

is used to remove some mistakenly detection areas. The threshold of the aspect ratio is set by an adaptive update method in order to obtain the most appropriate boundary condition for the detection video. Subsequently, the change rule of the face location between video frames is used to estimate the detection result in the current frame. Then, the estimate result is compared with the real detection result, and a contrast rule is used to modify the detection result of the deep learning network according to the contrast difference, deleting the false detection area, filling the missing detection area, which improves the detection accuracy.

The experimental results show that, no matter static single face detection or multiple faces detection under complicated conditions, besides the faster detection speed and stronger robustness against face rotation, the cascaded P-RBM learning network possesses the lower false detection rate and the lower missing detection rate. Moreover, combining it with the multi-inter-frame information fusion method to detect face in video not only keeps the faster detection speed and the lower false detection rate, but also reduces the missing detection rate significantly. In addition, it improves the detection performance of the partially obscured face.

Keywords： face detection in video, deep learning, Probability state-Restricted Boltzmann Machine(P-RBM), multi-inter-frame information fusion

目 录

第 1 章 绪论

1.1 研究背景及意义

近年来，视频网络覆盖程度越来越广，智能化的视频监控技术由于在国家安防、智慧城市等领域所具有的应用价值受到越来越多的关注。基于视频监控的人脸识别技术是目前身份识别技术的主流手段，相比于指纹识别、手指心电信号识别、DNA 识别等技术，它具有不易被观测对象发现的显著优势，从而能在目标不知情或不配合的情况下实现识别任务。人脸检测作为人脸识别的前提手段，是人脸识别系统的关键技术之一，其检测准确性对后续脸部特征提取及人脸识别的性能影响非常大。因此，人脸检测一直是人脸信息分析处理的一个研究热点。

人脸检测是指判断一幅经过一定处理和分析后的静止或动态图像中是否存在人脸，如果存在，则记录每个人脸的位置和大小信息[1, 2]，并在图像中标识出来。虽然人脸检测的研究已有几十年，并且也取得了一定的成果，但目前较成熟的人脸检测技术基本是针对一些理想情况下的检测，在实际应用中仍面临着许多挑战，其大致有以下几点：

1）人脸形状、表情的多样性。人脸有多种不同的形状，如方子脸、尖型脸、圆形脸等，且同一人脸形状在不同的人身上会有不同的大小比例。同时，人在不同情绪下会有不同的面部表情，开心时会笑，悲伤时会皱眉，而不同的面部表情会使面部器官发生不同的变化，如微笑会眯眼，大笑会张嘴。这些变化及人脸自身的多样性使得人脸很难用单一模板去匹配。

2）人脸旋转角度的多样性。人脸可以在水平方向和竖直方向上发生旋转，这使得图像中的人脸不是标准人脸（正面、双眼水平、鼻子竖直），部分人脸特征发生旋转或者消失。另外，由于摄像头一般都安置于较高处或较隐蔽处，导致拍摄到的人脸很少是与人视线平行的正面人脸。

3）饰物、遮蔽物的遮挡。视频人脸通常是在自然环境下拍摄的，因此人脸上可能会出现不同的饰物，如眼镜、帽子、胡须等，这会给人脸检测带来干扰。同时，视频中的人脸还会受到树叶、车辆、其余人脸等的遮挡导致部分脸部特征缺失。

4）复杂背景的影响。人脸检测的检测环境无法人为控制，在实际应用中极易出现与人脸形状、肤色相近的事物，这会造成很多误检。同时，不同的光照会导致不同的明亮度、对比度甚至造成阴影从而给检测带来难度。

5）设备的不同。不同的视频监控设备采集到的图像会在分辨率、色彩、大小等方面存在差异，并且不同的拍摄环境也会对采集设备有不同的要求，只有监控设备适应所监控的环境才能保证拍摄到高质量的图像，进而为后续检测提供保障。但是，在实际应用中，由于经济原因以及环境的多样性，很难保证所使用的监控设备是具有多种适应性的高质量图像采集设

备。

上述这些因素是视频人脸检测在实际应用中存在的困难与挑战，虽然现有的算法已经针对其中某种情况进行了改进，并取得了一些成果，但当存在上述多种情况时，这些算法仍不能带来较好的检测效果。因此，研究出一种能适应上述多种非理想情况下的视频人脸检测技术是当今的研究重点，这不仅具有很高的学术价值，也会给国家安防、社会安定等带来巨大贡献。

1.2 国内外研究现状

人脸检测作为人脸信息处理的首要步骤，是完成人脸识别、表情识别、人脸跟踪的前提手段。人脸检测开始于上世纪 60 至 70 年代[3]，作为人脸识别的一部分，其只需要对背景简单的正面人脸图像进行检测即可，因此这种简单的检测技术并未得到研究人员的重视。经过几十年的研究发展，理想环境下的人脸识别技术已经比较成熟。同时，随着数字通信、人机交互技术的不断发展以及国家安全和个人生命财产安全问题的日益突出，人脸识别系统开始在各种日常环境中使用，如商场、银行、海关、大型企业公司等。然而，原有的人脸检测技术已不能到达此时的应用要求，由此越来越多的研究者开始探索人脸检测问题。到了 20 世纪 90 年代初期，逐步发展起来的视频监控技术以及不断提高的计算机处理能力使得人们开始利用计算机实现半智能化的视频监控[4]。而在视频监控系统中，人脸信息具有极高的取证利用价值，同时监控人脸具有不易被观测者发现的显著优势，由此视频人脸检测应运而生。

视频人脸检测是为了确定视频监控范围内是否存在人脸，若存在则检测出其位置和大小。对于可视会议、视频监控、智能化设备以及智慧城市等方面，视频人脸检测都能提供很高的应用价值。

目前，越来越多的研究机构开始从事视频人脸检测的研究，在国外著名的研究机构有 MIT 的人工智能实验室和多媒体实验室[5]、CMU 机器人研究所[6]、ART 实验室以及美国国防部成立的 FERET 项目；国内知名的机构有清华大学计算机科学与技术系智能技术与系统国家重点实验室[7]、中国科学院计算机技术研究所[8]和中国科学自动化研究所等。随着对视频人脸检测研究的不断深入探索，关于视频人脸检测的文章层出不穷，与此相关的专题国际会议也不断召开，如 ICPR(IEEE International Conference on Pattern Recognition)[9]、ICIP(IEEE International Conference on Image Processing)[10]、CVRP(IEEE International Conference on Computer Vision and Pattern Recognition)[11]等。另外，越来越多的研究机构开始提供用于视频人脸检测算法研究的人脸库，常用的有：FERET 人脸数据库[12]，LFW 人脸数据库[13]、CAS-PEAL 人脸数据库[14]以及 PKU-SVD-B 人脸数据库等。大型人脸库的出现使得视频人脸检测算法的评估变得简便，同时各类算法间的比较也得以实现。到目前为止，虽然视频人脸检测算法多种多样，但根据实现思路，大致可以分为两大类：基于传统人脸检测方法的检测和基于连续性信息的检测。

1.2.1 基于传统方法的视频人脸检测

视频可以看成是多张图像按时间顺序连接而成，因此一类常用的视频人脸检测方法是将传统的人脸检测技术直接应用于单帧视频图像。传统的人脸检测方法有以下四类，分别为：基于先验知识的检测，基于模板匹配的检测，基于特征提取的检测以及基于机器学习的检测。

（1）基于先验知识的检测。

该方法利用人脸五官分布规则来实现人脸检测，主要依据人脸面部各器官之间的特定关系总结出一些描述人脸的规则，如人脸是对称的，鼻子位于中间，双眼关于中心对称等，再根据这些人为认知总结出来的规则对图像进行遍历搜索，将符合要求的区域作为人脸检测结果。

Reisfeld[15]等人首先根据广义对称变换理论和边缘图像寻找到人脸的对称轴，然后利用人脸五官分布所具有的限定条件以及眼睛和嘴巴位于对称轴上对称值最大处的特性定位人脸。Yang[16]等人提出利用一组等大小的马赛克块对人脸的五官区域进行分块，接着分别计算每个区域块的平均灰度，再根据一组规则获取人脸区域，最后利用边缘特征进行确认，这即为镶嵌图（又称为马赛克图）检测法。卢春雨[17]等人对上述这种方法进行优化，提出一种可以根据检测情况实时调整人脸划分为矩形区域时每个矩形块大小的检测方式，并且在计算每一个矩形块的平均灰度值时统计灰度及梯度的分布情况，继而进行人脸检测。为了能对人脸旋转具有一定鲁棒性，杨秋芬[18]等人在根据眼睛和嘴巴所构成的倒等腰三角形关系来检测正面人脸的同时，还利用侧面人脸中具有直角三角形结构关系的眼睛、嘴巴和耳朵来实现对旋转人脸的有效检测。

基于先验知识的方法实现比较简单，只需判断检测区域是否符合分布规则即可，但由于人脸本身的多样性以及人脸姿态、表情变化的多样性，要获得一个通用人脸五官分布规则是比较困难的，因此这类方法的漏检率和误检率较高。

（2）基于模板匹配的检测。

该方法通过计算待检测图像与事先设计的人脸模板之间的相似程度来检测人脸。早期的模板匹配法通常采用预定模板，如 Brunelli[19]等人提出一种用于解决正面人脸定位问题的形状模板匹配法，它首先利用 sobel 滤波器获取图像边缘特征，然后根据约束条件选取部分连通边界以确定头部轮廓，最后在头部区域内进一步匹配眼睛、嘴巴和眉毛区域从而实现检测；Leung[20]等人提出用对称分布的眼睛和鼻孔以及位于对称轴上的嘴巴这几个特征作为一个典型的人脸，之后用一系列高斯滤波器来提取各个方向和尺度上的特征，最后通过待检测人脸与每一个特征模板向量的匹配程度来确定人脸区域。

基于预定模板的匹配易于实现，但对例如尺寸、形状以及姿态等的变化适应性较差。因此，当检测环境较为复杂，人脸尺度、旋转角度都有变化时，基于可变形模板的匹配具有更好的检测效果。Yuille[21]等人首次提出一种可变形模板匹配法，它首先利用参数模板对弹性人脸特征进行建模，然后将待检测图像的边缘、波峰值、波谷值作为参数对模板进行实例化，

最后在其能量函数最小值处获取最优的弹性模板作为最佳匹配以实现对不同人脸的检测。尹雪聪[22]根据人脸不同部位在匹配过程中所起的作用不同赋予各个部件不同的权值，并且构建人脸各角度的可变形部件加权匹配模型，再将这些多角度模型进行融合以获取对各角度人脸较好的检测结果。虽然可变形模板对人脸各种变化的鲁棒性较强，但其能量函数中的各个权值只能通过经验来确定，并且优化能量函数的过程比较复杂，较难在实际应用中实现。

（3）基于特征提取的检测。

该方法利用脸部的不变性特征，如尺度特征、纹理特征、肤色特征等实现人脸检测。这类方法通过各种方式寻找人脸特征点，然后提取这些特征并利用这些特征构建统计模型以判断所检测区域是否为人脸。该方法不仅能通过存在的脸部特征检测人脸，还能通过它们相互之间的几何关系来检测人脸。

Anvar[23]等人提出一种基于 SIFT（Scale-invariant feature transform）特征的多人脸检测方法。首先提取待检测图像中的 SIFT 特征，接着计算这些特征与参考图集中各参考图的 SIFT 特征的欧式距离以筛选出有用特性，最后根据选取出的 SIFT 特征判定其所在区域是否为人脸区域。该方法的一大优势是会将检测出的人脸提取出来，作为参考图加入参考图集，用于之后的人脸检测，因此最初只需人工选取一张人脸参考图中的一对 SIFT 特征即可进行人脸检测，这使得整个检测过程基本无需人工参与。

肤色特征是如今最常用的特征提取对象，它仅受光照影响，因此对其的检测算法通常都具有较强的鲁棒性。基于肤色的人脸检测首先在特定颜色空间对肤色进行建模，然后通过该模型去判断待检测图像中的每个像素点是否为肤色点，最后再通过形态学处理获得人脸区域。常用的颜色空间有 RGB 颜色空间[24]、HSV 颜色空间[25]、YCbCr 颜色空间[26]等。肤色检测具有较快的检测速度，且能较好的去除背景区域，但是容易受到非人脸肤色区域和类肤色区域的干扰，因此目前的肤色检测常用于生成人脸候选区域以减少人脸检测范围。

基于人脸不变性特征的方法虽然对尺度、旋转等变化具有较好的鲁棒性，但是当其所利用的特征缺失或被噪声干扰时，检测效果将会受到很大影响。

（4）基于机器学习的检测。

这类方法依靠机器来学习人脸和非人脸样本，并利用统计分析来描述学习到的分类模型，再根据模式分类问题的解决思路对检测区域进行是否为人脸的分类，从而达到检测人脸的目的。相比于前面三种方法，这类方法不需要人为设定人脸特征的描述方式，而是由机器通过对大量样本的监督或非监督学习来获得具有最好分类性能的分类模型，因此其适应性较强。

人工神经网络是一类通过模拟人的大脑神经来进行数据处理分析的机器学习模型。Propp[27]等人通过构造一个具有 2 层隐藏层的 4 层神经网络来实现人脸检测，这是最早用于人脸检测的神经网络之一。此后提出的包括卷积神经网络（Convolutional Neural Network，CNN）、自组织映射神经网络（Self-Organizing Map，SOM 网络）、时延神经网络（Time Delay Neural Network ，TDNN）等都是通过对其结构进行优化得到的。Rowley[28]等人通过构建数个神经网络来实现对非正面人脸的有效检测，该方法通过对正面人脸和旋转人脸分别进行处理以达

到对于一定旋转角度人脸的较高检测准确率。

主分量分析法是一种基于线性子空间法的机器学习方式。它通过 K-L 变换寻找一个使各个分量相互独立的子空间，然后选取变换后的几个主要特征向量来描述输入图像，达到以最少特征向量描述输入数据最主要信息的目的，进而将输入数据的维数从像素个数减少为特征向量的维数，实现降维。Pentland[29]等人将主分量分析法引入到人脸检测中，首先对训练样本求取均值，接着计算它的协方差矩阵，并求得其对应特征值和特征向量，最后对这些特征向量进行正交化以选取能表示人脸的特征向量作为主分量构成投影矩阵，这些投影矩阵也称为“特征脸”。在检测时，首先将待检测图像转换到由这些特征向量构成的特征空间上，并计算投影后的图像与投影矩阵之间的距离，距离最小的即为人脸所在位置。

支持向量机（Support Vector Machine，SVM）是一种根据结构风险最小化原理在最小泛化误差处获得最优分类超平面，再利用该超平面实现模式分类的机器学习方式。Osuna[30]等人首先将 SVM 分类器应用于人脸检测，然后提出一种针对大规模数据集的 SVM 训练算法，使得基于 SVM 的人脸检测速度大大提高。

Adaboost 算法作为目前最常用的人脸检测算法，最早是由 Viola 和 Jones 提出的[31]。Adaboost 首先利用样本训练弱分类器，这些弱分类器只需比随机分类性能好即可，接着根据迭代的方法将若干个弱分类器组合成为强分类器，然后再以递增的原则构成级联分类器的每一层，以获得人脸检测需要的分类器。该算法利用训练好的级联分类器对待检测图像进行检测，在检测过程中不断丢弃被各分类器划分为非人脸的区域来减少进一步判断的区域个数，从而达到了减少计算量、提高检测效率的目的。

基于机器学习的人脸检测分类器在有限样本和计算能力的限制下，如果设计的相对简单，则有较好的鲁棒性，但分类正确率不高；如果采用复杂结构，可实现较高的正确率但鲁棒性较差。

视频人脸检测相比传统人脸检测，其难点在于如何在多变的检测背景条件下将图像中的脸部区域数据稳定的映射到语义人脸。因此仅依靠一种检测技术很难实现较好的检测性能。目前大多数将视频作为多帧静态图像处理的视频人脸检测技术都是结合上述两种或两种以上方法来实现的。黄禹馨[32]等人首先利用限定的肤色区域长宽比大小范围来去除背景区域，然后在剩余区域内利用灰度差分法和二值化法以确定嘴巴和眼睛的位置，最后根据眼睛和嘴巴所形成的倒三角关系来精确定位人脸。周瑾[33]等人先利用肤色检测获取肤色矩形区域，并构建该图像的肤色积分图，然后对肤色区域进行金字塔搜索，并在搜索过程中不断计算搜索窗口的肤色覆盖率，当覆盖率达到要求时便利用 Adaboost 算法对该搜索区域进行人脸检测，这种方式使得候选区域的选取更加准确。Sung[34]等人提出一种将模板匹配和神经网络分类器相结合的检测方法。首先根据正面人脸的分布特性建立模板，然后将人脸正样本和非人脸负样本与分布特性模板之间的匹配度作为训练数据训练神经网络分类器。在检测阶段，首先计算输入的待检测图片与人脸分布模板之间的差异度，并将结果作为输入数据传递给神经网络，最后利用神经网络对输入数据进行分类判断从而输出结果。另外，除了最初提出的基于 Haar

特征的 Adaboost 算法，研究人员已经将这种级联分类器的思想与其他特征相结合，如 Louis[35]等人将 LBP（Local Binary Patterns）特征与 Adaboost 的思想相结合，首先采用环形 LBP 特征级联成多级分类网络，利用环形 LBP 获取图像像素级特征，再在该分类网络的最后叠加一层基于 LBP 直方图特征的分类层，获取图像的块区域特征，进而较准确的实现视频人脸检测。

虽然通过结合多种人脸检测方式已经在一步步地改善视频人脸检测的检测效果，但上述这些基于单帧视频图像的检测所获取的所有信息仅来源于一张图像，而对于视频中多变的复杂背景，旋转人脸，以及部分被遮挡人脸，单帧图像无法提供足够的信息以应对这些非理想检测条件，因此越来越多的研究者开始着手利用视频所具有的连续性来实现具有更好检测性能的视频人脸检测。

1.2.2 基于连续性的视频人脸检测

相比于单张静态图像，视频中各帧图像之间具有连续性，它能为视频人脸检测提供额外的包括时空变换上的信息，因此越来越多的视频人脸检测开始根据这一特性利用运动目标检测技术来提取候选区域，然后在候选区域上进行人脸检测。运动目标检测根据背景是相对静止的，目标是相对运动的这一特点来实现，常用的方法有背景差分法、帧差法以及光流法。

（1）背景差分法。

该方法先对背景图像建模，然后逐像素比较检测图像与所建模型，像素差值非零的区域即为发生移动的区域[36]，即运动目标所在区域。谢仪[37]等人首先采用混合高斯模型来对背景图像进行建模，并采用在线 K-Means 算法对参数进行近似估计，以实现对复杂背景的及时更新，然后将待检测的监控视频图像与建模所获得的背景图像进行差分获得运动区域，最后利用 Adaboost 算法在运动区域上进行人脸检测。这种方法虽然能较快的去除背景区域，但当背景变化较多并且存在干扰的时候，所构建的背景模型并非为检测时的真实背景，从而导致获取的移动区域不是运动目标所在区域。

（2）帧差法。

这类方法通过相邻两帧或多帧视频图像中对应点的像素值变化程度，利用阈值判定来提取运动目标区域。向桂山[38]等人通过设置一个较大的阈值来保留下所有运动区域，再通过肤色以及一系列人脸特征规则进行人脸检测。但是这种基于相邻帧对应点像素值之差的帧差法极易受脉冲噪声的影响，因此 Xiong[39]等人提出一种基于块平均差的帧差法。它首先计算以某一像素点为中心的正方形区域内所有像素点的平均值，同时求取相邻帧对应于该点的平均像素值，通过两者之差与阈值的大小比较获得该点的二元像素值，之后用人脸轮廓模型获得二元图像中的人脸区域，最后用人脸的几何特征去准确定位人脸。帧差法是目前在运动目标检测中应用最多的方法，但是当目标运动的较为缓慢或者静止时，差分的效果不是很理想。

（3）光流法。

这类方法赋予速度矢量给图像中的每一个像素点，根据各像素点的速度矢量在后续过程中随时间变化的光流特性，区分运动目标与背景，从而实现运动目标的提取。龚卫国[40]等人

提出先提取图像角点，并通过 Adaboost 算法保留下人脸角点，再利用基于多分辨率的光流算法获得人脸角点的光流，最后根据所得的光流特性检测出 Adaboost 无法检测到的偏转人脸。但是这类方法在图像存在噪声或目标与背景相似的情况下并不能准确可靠的反应物体的运动信息。

另外，还有一类基于视频连续性信息的视频人脸检测方法是利用目标跟踪算法来实现的。Chang[41]等人提出一种在线自适应视频人脸检测算法，它首先通过 Adaboost 算法检测出各视频单帧图像中的人脸，然后利用视频所具有的连续性建立一种增量学习的高斯回归过程模型，该模型是通过对已有视频帧中的检测结果进行建模来获得当前检测视频帧中的人脸分布情况，以补上一些漏检的人脸，并且根据最新的人脸检测结果更新回归过程模型。但是这类方法只考虑了当前帧前面视频帧的信息，并没有考虑后面视频帧的信息。

总的来说，视频人脸检测是一个复杂的模式识别问题，即使结合上述多种方法也较难在非理想条件下达到一个很高的准确率，通常只能对于某一应用领域或特定情况才有较好的检测效果。因此针对能适应多种复杂情况，且稳定性高的视频人脸检测技术的研究是当前人脸检测和视频监控领域的热点。

1.3 本文的主要工作及章节安排

本文在系统的介绍已有的视频人脸检测相关研究的基础上，主要研究面向视频人脸检测的深度学习算法，它以现有的深度学习理论为基础，提出一个基于视频单帧人脸检测的深度学习检测网络，并结合视频帧间的连续性，实现具有较好检测性能的视频人脸检测。主要研究内容有以下几个方面：

1）论文针对非理想条件下快速准确的视频人脸检测问题，提出一个基于视频单帧的人脸检测算法。该算法首先对受限玻尔兹曼机进行改进，提出一种概率态受限玻尔兹曼机，用其概率表征来模拟人脑神经元连续分布的激活状态；然后以此为核心，通过级联多个概率态受限玻尔兹曼机来构造一个深度学习网络，利用该学习网络来模拟人脑对图像信息的处理过程，在充分提取输入数据各层次特征的基础上获得输入数据的语义信息，进而较准确的对输入数据进行是否为人脸的分类，实现对旋转角度、光照等变化具有较强鲁棒性的人脸检测。学习网络采用预训练来初始化网络参数，以逐层贪婪学习方法分层训练各概率态受限玻尔兹曼机，避免训练误差的多层传递，解决多层网络训练容易陷入局部最优的问题，并使得最后检测网络整体优化时有一个较好的初值。

2）论文提出一种多帧间信息融合的视频人脸检测算法。该算法在深度学习网络的检测基础之上，利用视频帧间的连续性，制定帧间信息传递规则，然后根据这些规则以及前后一帧或多帧视频图像的检测结果对当前帧检测结果进行修正。该算法根据视频内不同人脸与摄像头的距离基本相似的性质，通过人脸肤色区域长宽比允许范围去除一些非人脸区域。由于不同场景下人脸与摄像头的距离角度不同，长宽比也会有所不同，因此提出一种自适应的阈值更新方式以获取最适合当前检测环境的人脸肤色区域长宽比范围。其次，根据视频中的人脸

是从边缘逐渐移入，再从边缘逐渐移出的特性，利用相邻帧之间对应人脸的移动规律来估计当前帧该人脸的位置，并根据对比规则判断估计位置与检测位置之间的差异以修正部分漏检和误检。该方法相比常用的运动目标检测方法，是对已有检测结果的修正，同时也考虑了当前帧之后的视频帧信息，因此能获得更好的检测效果。

全文由五章内容组成，各章大致内容如下：

第一章，介绍视频人脸检测的研究背景及意义，并简单介绍了国内外研究现状，主要的两类视频人脸检测方法及其优缺点。

第二章，详细介绍深度学习的原理、优势、与神经网络的关系，并简要介绍了目前深度学习的两种典型网络模型及其应用。

第三章，提出基于视频单帧的人脸检测算法。首先提出概率态受限玻尔兹曼机，利用其神经元的概率表征来模拟人脑神经元从最活跃到最不活跃的连续分布激活状态，然后级联多个概率态受限玻尔兹曼机以仿真人脑对视觉的分层次学习模式，最后通过对肤色检测所生成的候选区域进行是否为人脸的分类来完成人脸检测。该学习网络以逐层递减隐藏层神经元数来控制网络规模，采用分层训练和整体优化的机制来缓解鲁棒性和准确性的矛盾。

第四章，提出多帧间信息融合的视频人脸检测算法。该算法利用连续视频帧之间人脸在肤色区域长宽比和位置信息上的相关性，制定人脸区域删补规则，从而对深度学习网络的检测结果进行修正，删去误检区域，补上漏检区域，达到降低漏检率和误检率的目的。

第五章，对本文工作进行总结，提出不足之处和今后的研究发展方向。

第 2 章 神经网络与深度学习

2.1 概述

人工神经网络是一种模拟人脑智能特点和信息处理机制的机器学习网络，简称为神经网络。它通过模拟人脑神经细胞的工作特点，即单元间的广泛连接、并行分布式的信息存储与处理、自适应的学习能力等，使机器具有人脑那样的感知、学习和推理功能。

人工神经网络的特性如下：

1）信息在神经网络中是分布式存储的。特定信息的表示是通过神经网络中神经元之间的连接关系及强度实现的。

2）信息在神经网络中的处理是并行执行的。每个神经元互不干扰地运算和处理接收到的信息。

3）自组织性和自学习性是神经网络对信息的特有处理过程。神经网络中的权值可以自适应的变化，这一变化过程就是神经元的学习过程。

4）容错性较强。神经网络能进行特征提取、聚类分析、缺损模式复原等工作，因此能够识别带有一定噪声或畸变的输入模式。

基于以上这些特点，人工神经网络自提出以来就被广泛研究，至今已有感知器、后向传播网络、自组织特征映射神经网络、Hopfield 网络等多种神经网络模型。但是这些神经网络在实际应用中都采用浅层结构。所谓浅层结构，就是指该神经网络只含一层隐藏层，它只能对输入数据进行一次映射，使其转换成特定映射空间特征的简单结构，而仅有的一次映射操作并不能提取足够的有用信息。虽然上述这些神经网络也曾试图建立多隐藏层结构，但由于训练上的难度这一想法一直未能获得很好的结果。与此同时，神经科学研究发现，识别技术中的信号预处理过程并不真实存在于哺乳动物执行识别任务的过程中，大脑执行识别的真正过程是在其复杂层次结构中传播输入信号，通过每一层次对输入信号的重新提取和表达，最终让哺乳动物感知世界[42, 43]。这一发现进一步激发人们对多层神经网络的探究。

2006 年 Hinton[44]等人提出多隐藏层人工神经网络具有较强的特征学习能力，学习到的特征对数据有更加本质的刻画，而多层神经网络在训练上的难度可以通过逐层训练来有效克服，也就是所谓的深度学习（Deep Learning）。它利用计算机模拟人脑皮层的信号处理机制，通过多层分段式学习，形成逐段递进的非线性映射关系进而实现感知视觉的功能。它在提取数据表象特征的基础上逐层组合抽象获得输入数据各层次特征并最终获得语义特征，通过不断对训练样本语义特征地学习实现对能表征输入数据特性的复杂函数逼近，从而较准确地实现识别、分类任务。相比浅层结构，深度学习所具有的深层结构更加突出了特征学习的重要性，即利用多隐藏层结构实现特征的逐层变化，进而将输入数据逐层映射到不同特征空间以提取

不同层次特征，更好更全面的获得了输入数据想要表达的信息。

目前大多数提到深度学习的地方都是指多隐藏层神经网络，但不能将深度学习与多隐藏层神经网络对等，深度学习的理念更加广泛，它是指一种多层次、非线性的学习过程。

2.2 深度学习研究现状

深度学习是机器学习的一个新的研究阶段，它通过建立类似于人脑分层次学习的机制，对输入数据逐级提取从底层到高层的特征，从而建立从底层信号到高层语义的映射关系[45]。同时，深度学习通过逐层构建的多层网络使其能学习隐藏在数据内部的关系，从而使学习到的特征更加具有推广性和表达力[43]。对于识别任务，人脑并不会对获得的原始信息进行某一特定的预处理，而是直接将其在大脑复杂的层次结构中传播，通过每一层对输入信号进行重新提取和表达最终让人脑识别出输入信号想要传达的信息[42, 46]。这种分层次处理信息、提取特征的过程正是深度学习网络模型的构建基础。本节主要介绍两种典型的深度学习网络模型：深度置信网络和卷积神经网络，以进一步阐述深度学习的实现过程，并对其最新的应用成果进行简要介绍。

2.2.1 深度置信网络

深度置信网络（Deep Belief Network，DBN）是 Hinton[47]提出的有向图模型，其结构如图 2.1 所示

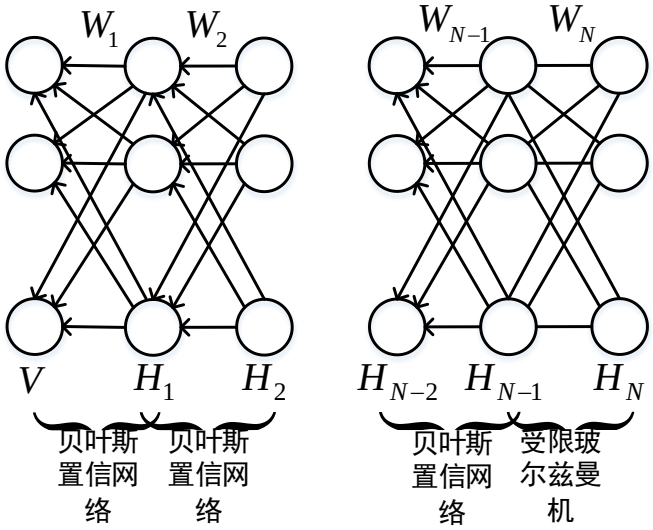


图 2.1 深度置信网络

它由一层可视层V，多层隐藏层$H_i(\mathrm{i}=1,2,\mathrm{L},N)$组成。DBN 中最右边两层构成受限玻尔兹曼机（Restricted Boltzmann Machine，RBM），其余相邻层构成贝叶斯置信网络，置信函数选择式（2.1）的 sigmoid 函数：

$$sigmoid(x)=\frac{1}{1+e^{-x}} \tag{2.1}$$

贝叶斯置信网络是一种后验概率分布模型，其后验概率为

$$P(H_i \mid H_{i-1}) = \frac{P(H_{i-1} \mid H_i)P(H_i)}{P(H_{i-1})} \tag{2.2}$$

其中 $i = 1,2,\mathrm{L}\ ,N$（为了统一，这里用 H_0 表示 V ）；$P(H_{i-1} \mid H_i) = sigmoid(C + WH_i)$，$C$ 为 H_{i-1} 的偏置，W 为 H_{i-1} 和 H_i 之间的连接权值；$P(H_{i-1})$ 由输入或者前一层的输出给定；$P(H_i)$ 为已知结果。在 H_N 层某一神经元被激活，其余神经元未激活的条件下，DBN 利用贝叶斯网络将 H_N 中的数据信息传递到可视层，通过可视层获得的数据信息来判断该激活神经元所表示的特征。

DBN 是一个概率生成模型，它不同于传统判别模型的分类网络只在观察数据下对数据标签的条件分布进行估计，而是会建立一个观察数据和数据标签之间的联合分布，以获得观察数据和数据标签相互间的条件分布，即 $P(data \mid label)$ 和 $P(label \mid data)$。DBN 的联合概率分布为：

$$P(V, H_1, H_2, \mathrm{L}\ , H_N) = P(V \mid H_1)P(H_1 \mid H_2)\mathrm{L}\ P(H_{N-2} \mid H_{N-1})P(H_{N-1}, H_N) \tag{2.3}$$

对于 DBN 的训练，Hinton 提出首先利用预训练[44]来初始化网络参数，并在其中采用逐层贪婪学习算法[47]将多层网络分层训练，这不仅避免由于训练误差多层传递使得最后一层获得的误差太小以至于不能很好地调整网络参数的情况，同时保证后续的整体优化过程能有一个较好的初始值，解决多层网络训练容易陷入局部最优的问题。更加重要的是，这种多层网络分层训练的方式能够提高分类正确率。预训练的具体过程就是将 DBN 内相邻的两层（如 H_{i-1} 和 H_i ）看成一个只有一层输入层（ H_{i-1} ）和一层隐藏层（ H_i ）的完整神经网络，采用该神经网络的训练方式进行训练。由于该神经网络只有一层隐藏层，所以结构简单，对随机初始化的网络参数进行训练就能达到一个较好状态。之后将该隐藏层（ H_i ）与下一隐藏层（ H_{i+1} ）相连，构成新的神经网络，并且将该隐藏层（ H_i ）在前一次训练所得参数下获得的数据作为此次新的神经网络的输入数据进行训练，按此训练方式一直训练完 DBN 的所有隐藏层从而完成预训练。预训练获得的是单个子网络的最优参数，因此还需要一个整体优化过程将单个子网络的最优参数优化为整个 DBN 的最优参数，通常采用后向传播算法[48]或 wake-sleep 算法[49]来实现。这种多层网络分层训练的思想已经成为目前深度学习网络的主要训练方式，不管采用何种神经元，何种子网络类型，运用这种训练方式都能较好的解决多隐藏层神经网络在训练上的困难，使得深度学习网络的应用成为可能。

除了基于受限玻尔兹曼机和贝叶斯置信网络的深度学习网络，比较常用的还有基于自动编码器、玻尔兹曼机、径向基函数的深度学习网络，这些学习网络和 DBN 相比主要差别仅在于子网络类型不同，而数据传递的思想、实现原理、训练过程都是一致的，因此本节只介绍了 DBN。

目前，DBN 及其改进模型已被越来越多的用于解决模式识别问题。Luo[50]等提出一种可切换深度学习网络用于行人检测，首先构建多个 DBN 来对应人体的局部特征，如肩膀、头部、

腿部等，然后利用一个可切换变量将各个 DBN 连接进一个分类层，可切换变量决定其对应 DBN 的输出结果是否在分类层中起作用，即选择出最适合输入数据的各局部特征组合方式，最后根据组合结果进行分类。Nakashika[51]等提出一种考虑数据间时序依赖性的 DBN 用于语音识别。该学习网络中的隐藏层不仅接收前一隐藏层的输出数据，还直接接收可视层的数据，通过两者共同的作用来提取语音序列中的特征。但是，该方法是通过对演讲者和测试者分别建立 DBN 网络模型，然后比较两者在输出上的差异进行识别，因此只适合检测对象较少的情况。

2.2.2 卷积神经网络

不同于 DBN 的子网络级联结构，卷积神经网络（Convolutional Neural Network，CNN）是一种权值共享模型，其结构如图 2.2 所示。

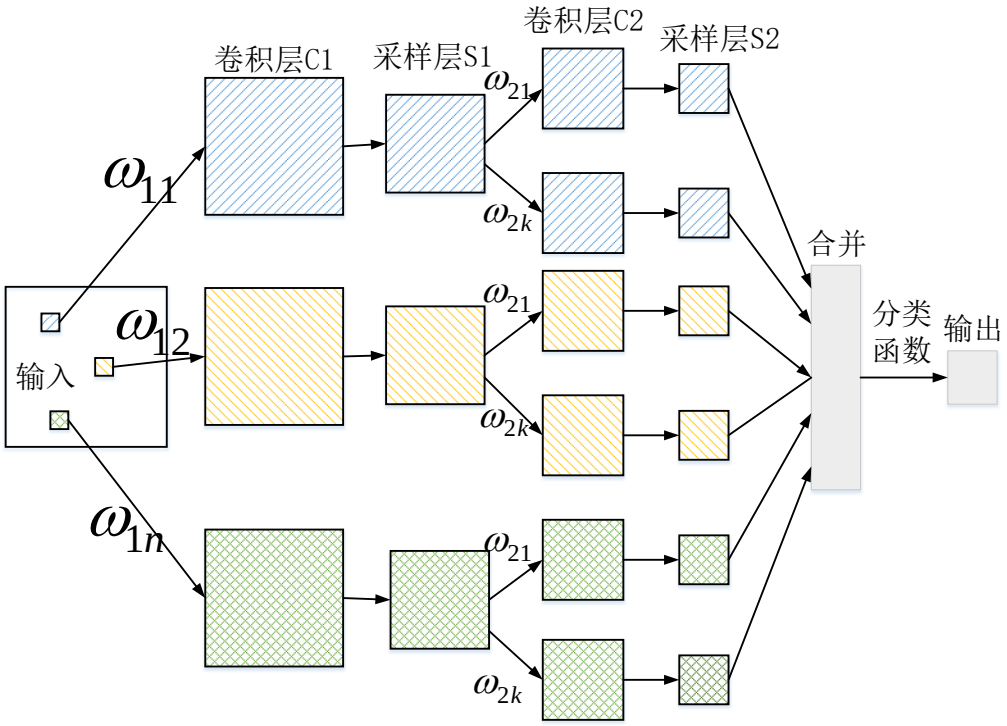


图 2.2 卷积神经网络

它由多个相邻的卷积层 C_i 和采样层 S_i 组成（图 2.2 中仅画了两层），每一层都是由多个二维平面组成。其中，卷积层是由多个不同的卷积核与前一层每一个二维平面卷积得到。因此，当前一层有 n 个二维平面，该层有 k 个卷积核时，该卷积层一共有 $n\times k$ 个二维平面。卷积过程如下：

$$x_j^{Cl} = f(x_{\lceil j/k \rceil}^{Sl-1} * \omega_{li} + b_j^{Cl}) \tag{2.4}$$

其中 x_j^{Cl} 表示第 l 卷积层中第 j 个二维平面；$x_{\lceil j/k \rceil}^{Sl-1}$ 表示第 $l-1$ 采样层中第 $\lceil j/k \rceil$ 个二维平面，$\lceil j/k \rceil$ 表示取大于或等于 j/k 的第一个整数，k 表示该卷积层一共有 k 个卷积核；ω_{li} 表示第 l 卷积层中第 i 个卷积核；b_j^{Cl} 表示第 l 卷积层中第 j 个二维平面的偏置；$*$ 表示卷积操作；$f(\cdot)$ 是一个映射函数，常用的有 sigmoid 函数、径向基函数、修正的 sigmoid 函数 $abs(g\times\tanh())$ 等。

卷积核每次只作用于输入平面的局部区域，称该小区域为局部感受区，该感受区经过卷积映射操作后产生的值赋给卷积层中对应二维平面上对应位置的神经元，因此该神经元的值保留了该感受区的某一特征属性，而该神经元的位置保留了感受区的位置信息。不同的卷积核用于提取不同的特征信息，同一卷积核在二维平面中滑动以判断不同位置上是否具有该卷积核所表示的特征。

采样层是对卷积层中每一个二维平面上的每一个 $m\times m$ 小区域进行采样，再经过函数映射得到的，即

$$x_j^{Sl} = g(\beta_j^{Sl} sample(x_j^{Cl}) + b_j^{Sl}) \qquad (2.5)$$

其中 x_j^{Sl} 表示第 l 采样层中第 j 个二维平面； x_j^{Cl} 表示第 l 卷积层中第 j 个二维平面； $sample(\cdot)$ 表示采样函数，通常为 $m\times m$ 区域内所有像素值的和； β_j^{Sl} 表示第 l 采样层中第 j 个二维平面的乘性偏置； b_j^{Sl} 表示第 l 采样层中第 j 个二维平面的加性偏置； $g(\cdot)$ 是一个映射函数。

因此卷积层有 $n\times k$ 个二维平面，采样层就有 $n\times k$ 个二维平面，但尺度明显减小，这使得采样后的数据保留下有用信息，去除了冗余信息。最后一层采样层的各平面数据会被重排成一维向量，然后由一个分类函数对其进行分类。

CNN 是采用有监督的训练方式，利用极小化实际输出与理想输出的差以反向调整卷积网络的参数。由于每个卷积核所表示的特征不是人为设定，而是通过对样本的反复训练得到的，这使得各个卷积核所学习到的特征具有较强的鲁棒性。同时，CNN 利用不同的卷积核依次作用于图像的局部区域，使得图像的不同区域共享各个卷积核；每个卷积核会产生相应的卷积平面，通过采样来保留卷积层中各平面内的有用信息，去除冗余信息，降低复杂度。CNN 这种局部权值共享的特殊性不仅降低了神经网络的复杂度，同时使其对尺度缩放、平行移动、旋转等各种形式变化具有较高的不变性，而不同卷积核之间相互无影响的特性使得 CNN 得以并行训练。另外，卷积层和采样层相邻的设计模式也更接近于真实的生物神经结构，将其应用于多维图像可以避免在特征提取和分类过程中由于数据重建而引起的复杂度。

CNN 的这种特殊结构使其在语音识别和图像处理方面有着独特的优越性。Krizhevsky[52] 等首先将 CNN 应用于图像分类，提出一个具有五层卷积层，共 650000 个神经元的 CNN 网络，在 ILSVRC-2012 数据集上将误检率降到了 16.6%以下。程文博[53]等将卷积神经网络应用到注塑制品短射缺陷识别当中，提出一种具有两层卷积层、两层采样层、一层感知层以及一层输出层的识别网络模型，克服了现有缺陷识别需要人工提取缺陷特征的不足，并且将识别率提高到 99%。Sun[54]等首先构建 60 个具有四层卷积层、三层采样层、一层分类层的 CNN 用于提取人脸不同区域的各种特征，然后利用一个联合贝叶斯网络将 60 个 CNN 所提取的特征进行融合来实现人脸认证，将人脸识别率提高到了 97.46%。

2.3 本章小结

本章首先对神经网络进行简要介绍，从其不足之处引出深度学习这一概念。通过深度学习的由来、原理及优势等方面详细介绍了深度学习理论。同时，本章介绍了两种典型的深度

学习网络模型：深度置信网络和卷积神经网络，两者的主要区别在于深度置信网络是由多个简单神经网络级联而成，它利用各隐藏层提取输入数据各层次特征，是各种级联型深度学习网络的代表；而卷积神经网络是通过共享卷积核的方式将不同卷积核依次作用于图像的不同区域来提取不同位置上的不同特征。本章对于深度学习的研究为后续章节奠定了理论基础。

第 3 章 基于视频单帧的人脸检测算法

3.1 引言

无论是在日常生活还是安防领域，人脸作为监控的主要目标具有较高的取证利用价值。同时，随着视频监控设备的普及，具有较高准确率和较快检测速度的视频人脸检测技术成为了当今的研究热点。

面向视频单帧人脸检测的深度学习算法就是针对上述这一研究热点而提出的。深度学习是对人脑分层次学习过程的模拟，由于人的视觉系统对图像信息的处理是分级的，它是一个从边缘特征到轮廓特征、从局部特征到全局特征、从整体信息到核心信息、从数据信息进而产生智慧语义的过程，人脑学习从图像中分辨出人脸的过程与此相同。本章结合这种分级处理机制，提出一种由一层可视层、四层隐藏层以及一层分类层组成的深度学习网络，从而模拟人脑对图像信息的处理过程以提取输入数据各层次特征，并建立各层次间的映射，学习隐藏在数据内部的关系，实现对复杂背景、人脸变化、光照等因素具有较强鲁棒性的人脸检测。

本章首先简要介绍受限玻尔兹曼机，然后根据人脑神经元的真实状况对其进行改进提出一种概率态受限玻尔兹曼机（Probabilistic state-Restricted Boltzmann Machine，P-RBM），并以此为核心，级联多个 P-RBM 构造面向视频单帧人脸检测的深度学习网络。该学习网络在充分提取输入数据各层次特征的基础上获得输入数据的语义信息，从而能较好的实现非理想情况下的人脸检测。另外，为了加快检测速度，本章先将检测图像通过肤色检测生成候选区域以缩小人脸检测范围，然后再利用该学习网络实现对候选区域的人脸和非人脸分类，最后在图像中标定人脸完成检测。对于学习网络的训练，通过预训练[44]来初始化网络参数，并在预训练过程中采用逐层贪婪学习[47]将多层 P-RBM 分层训练，从而在避免训练误差多层传递的同时使得后续的整体优化过程能有一个较好的初始值，不仅解决了多层网络训练可能陷入局部最优的问题，还进一步提高了检测准确率。

3.2 受限玻尔兹曼机及其改进

本节先简要介绍受限玻尔兹曼机的原理及其模型，然后通过分析其与人脑神经元真实状态间的差异对其进行改进，提出一种概率态受限玻尔兹曼机，它继承了受限玻尔兹曼机的优点并对其不足之处进行完善，从而能更好的完成学习分类任务。

3.2.1 受限玻尔兹曼机

受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）是 Smolensky[55]通过限制玻尔兹曼机中同一层神经元彼此之间无连接而提出的一种随机神经网络模型。随机神经网络是指网络中的神经元是随机神经元，神经元的状态只有激活和未激活两种，通常采用二进制中的 1 来表示激活状态，0 来表示未激活状态，由神经元的概率与一个 0~1 之间的随机数进行大小

比较来决定其状态。RBM 由一层可视层V和一层隐藏层H组成，结构如图 3.1 所示。其中$v_i(i=1,2,\mathrm{L},M)$表示可视层第i个神经元，$h_j(j=1,2,\mathrm{L},N)$表示隐藏层第j个神经元。

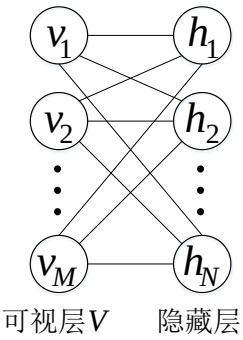


图 3.1 受限玻尔兹曼机结构示意图

RBM 是一个基于能量函数的神经网络，它能够描述变量之间的高阶相互关系。能量函数是描述整个系统的一种测度，系统越有序或概率分布越集中，系统的能量就越小，因此能量函数的最小值对应系统的最稳定状态。RBM 的能量函数定义如下：

$$E(V,H)=-C^TV-B^TH-V^TWH \tag{3.1}$$

其中，C表示可视层的偏置，B表示隐藏层的偏置，W表示可视层和隐藏层之间的连接权值。

基于能量函数的模型的联合概率分布如下：

$$P(V,H)=\frac{1}{Z}e^{-E(V,H)} \tag{3.2}$$

其中$Z=\sum_{V,H}e^{-E(V,H)}$称为配分函数，用于归一化。因此，RBM 的联合概率分布为：

$$P(V,H)=\frac{e^{-E(V,H)}}{\sum_{V,H}e^{-E(V,H)}}=\frac{e^{C^TV+B^TH+V^TWH}}{\sum_{V,H}e^{C^TV+B^TH+V^TWH}} \tag{3.3}$$

由此可以得出 RBM 中可视层V和隐藏层H的条件分布为：

$$P(H|V)=\frac{P(V,H)}{\sum_{H}P(V,H)}=\prod_j P(h_j|V) \tag{3.4}$$

$$P(V|H)=\frac{P(V,H)}{\sum_{V}P(V,H)}=\prod_i P(v_i|H) \tag{3.5}$$

从式（3.4）和式（3.5）可以得出，RBM 中同一层神经元彼此之间是条件独立的，这一性质使得求解 RBM 中各神经元的概率变得简便，极大地简化了神经网络的复杂度，加快了运行速度。

3.2.2 概率态受限玻尔兹曼机

RBM 所具有的对变量间高阶关系的描述能力，以及同一层神经元彼此条件独立的性质使

其成为了目前神经网络的主要结构。然而其神经元只有激活和未激活两种状态的性质与人脑神经元的真实状况并不十分相符。虽然人脑神经元的状态只有兴奋和抑制两种，但其输出或传达的信号并非是简单的二元逻辑值。当神经元被激活时，通常是发出一个类似于调频信号的一串脉冲，该脉冲信号的密度是可以表达连续量的[56]。同时，乔晓艳[57]等对大脑神经元兴奋性改变的实验也进一步证明大脑神经元的兴奋状态并非是 0 与 1 之间的突变，而是一个类似概率分布的连续变化。因此，这里对 RBM 中的神经元状态进行改进，在获得其神经元的激活概率后，不再将其变为二元逻辑值，而是直接用其概率来模拟人脑神经元连续分布的激活状态，称之为概率态。采用概率值而非二元值不仅能降低前一层网络的误判对后续网络的影响，提高神经网络的检测准确率，还能解决因为对一些相对弱小信息的屏蔽导致学习过程陷入局部最优的问题。我们称改进后的这一神经网络为概率态受限玻尔兹曼机（Probability state-Restricted Boltzmann Machine， P-RBM）。

由式（3.3）、式（3.4）可以推出，当给定可视层时隐藏层中某一神经元被激活的条件概率为：

$$\begin{aligned} P(h_j=1|V) &= \frac{e^{C^TV+b_jh_j+V^Tw_jh_j}}{\sum_{h_j} e^{C^TV+b_jh_j+V^Tw_jh_j}} \\ &= \frac{e^{b_jh_j+V^Tw_jh_j}}{\sum_{h_j} e^{b_jh_j+V^Tw_jh_j}} \\ &= \frac{e^{b_j+V^Tw_j}}{1+e^{b_j+V^Tw_j}} \\ &= \frac{1}{1+e^{-(b_j+V^Tw_j)}} = sigmoid(b_j+V^Tw_j) \end{aligned} \tag{3.6}$$

其中 b_j 表示隐藏层中第 j 个神经元的偏置， w_j 表示权值 W 的第 j 列。

同理可得，当给定隐藏层时，可视层中某一神经元被激活的条件概率为：

$$P(v_i=1|H) = \frac{1}{1+e^{-(c_i+w_iH)}} = sigmoid(c_i+w_iH) \tag{3.7}$$

其中 c_i 表示可视层中第 i 个神经元的偏置， w_i 表示权值 W 的第 i 行。

3.3 级联型 P-RBM 深度学习检测网络

依据人的视觉系统对图像信息从边缘特征到轮廓特征再到局部特征最后到整体特征进而产生智慧语义的分层处理过程，提出一种以 P-RBM 为核心，由一层可视层、四层隐藏层，以及一层分类层组成的级联型 P-RBM 深度学习检测网络，网络中的可视层和四层隐藏层构成了四个 P-RBM，对应上述视觉系统对图像信息的四层处理过程。另外，为了加快检测速度，在预处理层利用肤色检测生成候选区域作为该检测网络的输入，再根据检测网络的输出结果在图像中标定人脸完成检测。

检测网络模型如图 3.2 所示，V 表示可视层，$H_i(i=1,2,3,4)$ 表示第 i 隐藏层，$W_i(i=1,2,3,4,class)$ 表示相邻层之间的连接权值。图中各层纵向排列，其长度对应该层神经元数，层内神经元彼此之间无连接；隐藏层 H_1 的神经元个数通常大于可视层 V，这有利于充分提取输入数据最底层特征；其后各隐藏层（H_2、H_3、H_4）神经元数逐层递减以去除冗余信息并缩减网络规模提高检测速度；可视层与第一隐藏层以及相邻隐藏层所构成的四个 P-RBM 组成了级联型 P-RBM。这种结构设计的目的在于既能够让神经元提取足够的底层信息以生成高层语义，又能够约束网络中神经元的规模以降低局部最优的困扰，同时还能提高计算效率。最后由分类层对其获取的语义信息进行判定以输出检测结果。另外，这里将候选区域数据转换为一列的一维数据进行输入，方便计算。

级联型P-RBM

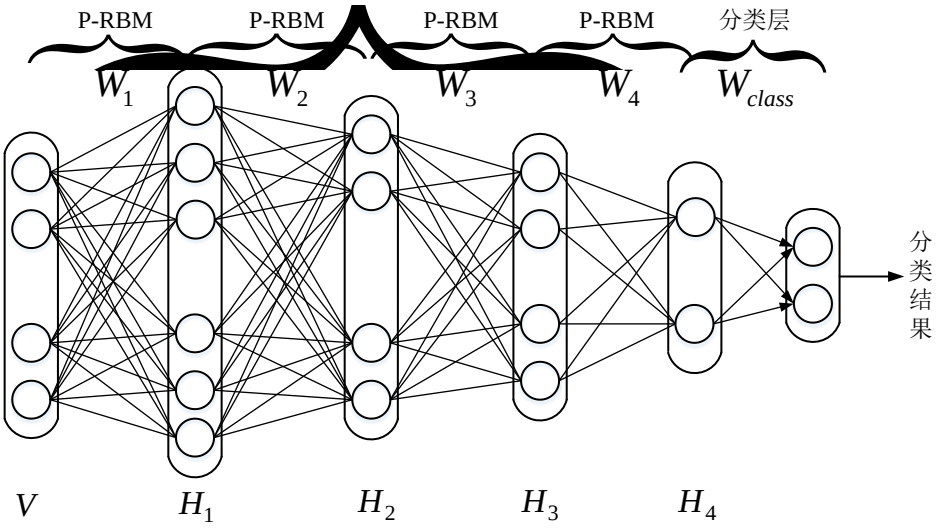


图 3.2 级联型 P-RBM 深度学习检测网络

由于直接训练多层神经网络需要将误差从分类层传递到可视层，多层传递使得可视层接收到的误差很小，不仅不利于调整网络参数，还容易导致神经网络陷入局部最优，因此这里将检测网络的训练分为三个阶段：级联型 P-RBM 训练，分类层训练以及整体优化。级联型 P-RBM 训练采用无监督学习方式使检测网络充分学习输入数据各层次特征，它构成了整体优化的预训练过程；分类层训练是在级联型 P-RBM 训练所得的基础上根据 H_4 的输出采用有监督学习方式来保证训练的准确性；整体优化将前两阶段所得结果作为初始值，对检测网络进行总体调整。

3.3.1 级联型 P-RBM 训练

级联型 P-RBM 训练以逐层贪婪学习方法[47]将多层 P-RBM 分层训练，避免训练误差的多层传递。P-RBM 训练的目的是通过参数的调整来学习输入数据的概率分布模型。其输入层（即可视层）的概率密度函数为：

$$p(V)=\sum_{H}\frac{e^{C^TV+B^TH+V^TWH}}{\sum_{V,H}e^{C^TV+B^TH+V^TWH}} \tag{3.8}$$

其中C表示可视层的偏置；B表示隐藏层的偏置；W表示可视层与隐藏层之间的连接权值。通过对可视层概率密度函数的对数似然函数求导可得：

$$\begin{aligned}\frac{\partial \ln p(V)}{\partial w_{ij}}&=p(h_j=1|V)\,p(v_i=1|H)-\sum_{V}p(V)p(h_j=1|V)\,p(v_i=1|H)\\ \frac{\partial \ln p(V)}{\partial b_j}&=p(h_j=1|V)-\sum_{V}p(V)p(h_j=1|V)\\ \frac{\partial \ln p(V)}{\partial c_i}&=p(v_i=1|H)-\sum_{V}p(V)p(v_i=1|H)\end{aligned} \tag{3.9}$$

式（3.9）中的$\sum_V P(V)$项需要通过多次吉布斯采样[58]在获得输入数据概率密度函数的条件下对输入数据所对应的概率值求和来获得，但该方法计算量大。这里采用文献[59]所提的对比差异算法（Contrastive Divergence，CD）[59]来减少计算量。K 阶对比差异的定义式如下：

$$CD-k=P_0\,\|\,P_\infty-P_k\,\|\,P_\infty \tag{3.10}$$

其中P_0表示初始化网络参数时的概率分布；P_∞表示输入数据确切的概率分布；P_k表示k次吉布斯采样之后得到的概率分布；$P_i\,\|\,P_\infty$表示P_i与P_∞之间的 KL 距离。由于每一次吉布斯采样之后获得的概率分布都会比前一次更接近于输入数据确切的概率分布，因此

$$P_{k-1}\,\|\,P_\infty >= P_k\,\|\,P_\infty \tag{3.11}$$

等号仅在P_{k-1}与P_k为相同的概率分布时成立，此时的概率分布即为输入数据确切的概率分布。因此一般设置$k=1$，从而在其最小值处$(CD-1=0)$获得P_0与P_1所满足的相同概率分布。同时，仅有一次的吉布斯采样又能在很大程度上简化 P-RBM 训练的复杂度，加快训练速度。

其中，一次完整的吉布斯采样过程包含以下两步：

1）对隐藏层中每一个神经元h_j采样获得其在可视层和该隐藏层其余神经元状态已知条件下的概率$p(h_j\,|\,H_j\},V)$，其中$H_j\}$为该隐藏层除h_j外的其余神经元。由于 P-RBM 中隐藏层各神经元彼此之间是条件独立的，因此

$$p(h_j\,|\,H_j\},V)=p(h_j\,|\,V) \tag{3.12}$$

2）根据第 1 步获得的隐藏层状态，采样重构出可视层各神经元的概率状态$p(v_i\,|\,V_i\},H)$，其中$V_i\}$是除所求神经元v_i外可视层的其余神经元。由于 P-RBM 中可视层各神经元彼此之间也是条件独立的，因此

$$p(v_i\,|\,V_i\},H)=p(v_i\,|\,H) \tag{3.13}$$

由一次完整的吉布斯采样，可以获得可视层各神经元概率，然后通过概率建模获得吉布

斯采样后可视层的概率分布 P_1，继而可求得 $CD-1$ 在各参数下的偏导，如下：

$$\begin{aligned}
&\frac{\partial}{\partial w_{ij}}(CD-1)=<p(v_i=1|H)p(h_j=1|V)>_{P_0}-<p(v_i=1|H)p(h_j=1|V)>_{P_1}\\
&\frac{\partial}{\partial b_j}(CD-1)=<p(h_j=1|V)>_{P_0}-<p(h_j=1|V)>_{P_1}\\
&\frac{\partial}{\partial c_i}(CD-1)=<p(v_i=1|H)>_{P_0}-<p(v_i=1|H)>_{P_1}
\end{aligned} \tag{3.14}$$

特别需要注意的是，式（3.9）和式（3.14）中的 $p(v_i=1|H),\ p(h_j=1|V)$ 即为 P-RBM 中神经元的概率态，其中 $p(v_i=1|H)$ 的初始值由输入给定。

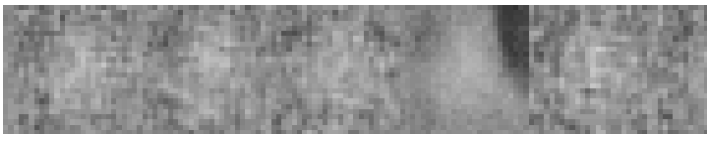
由式（3.14）可得 P-RBM 中各参数的更新公式如下：

$$\begin{aligned}
&\Delta w_{ij}=momentum\cdot\Delta w_{ij}+\varepsilon(<p(v_i=1|H)p(h_j=1|V)>_{P_0}-\\
&\qquad\qquad <p(v_i=1|H)p(h_j=1|V)>_{P_1}-w_{cost}\cdot w_{ij})\\
&w_{ij}=w_{ij}+\Delta w_{ij}\\
&\Delta b_j=momentum\cdot\Delta b_j+\varepsilon(<p(h_j=1|V)>_{P_0}-<p(h_j=1|V)>_{P_1})\\
&b_j=b_j+\Delta b_j\\
&\Delta c_i=momentum\cdot\Delta c_i+\varepsilon(<p(v_i=1|H)>_{P_0}-<p(v_i=1|H)>_{P_1})\\
&c_i=c_i+\Delta c_i
\end{aligned} \tag{3.15}$$

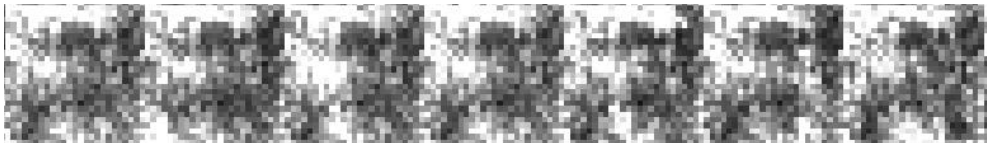
其中 Δw_{ij}，Δb_j，Δc_i 分别表示权值和偏置的更新量；$<\cdot>_{P_i}$ 表示数据在 $P_i(i=0,1)$ 下的均值；ε 表示学习率，是一个比例因子；$momentum$ 表示动量因子，它在此次变化量上加上一个正比于前次变化量的值 $momentum\cdot\Delta\theta$（其中 $\theta=w_{ij},b_j,c_i$），使 P-RBM 能滑过局部最小值；w_{cost} 表示权重衰减因子，用于减少之前训练得到的权值对后续权值更新的影响[60]。

利用式（3.15）不断更新参数，在获得满足 $CD-1=0$ 的参数值时完成 P-RBM 的训练，将该 P-RBM 隐藏层的数据作为下一个 P-RBM 可视层的初始数据重复上述训练步骤，直至完成所有 P-RBM 的训练。这种多层网络分层训练并在单层网络训练中运用对比差异算法[59]调整参数的训练方式相比传统的随机初始化多层网络参数然后利用误差反向传播来优化参数的训练方式，避免了误差的多层反向传播，解决了多层网络训练容易陷入局部最优的问题。

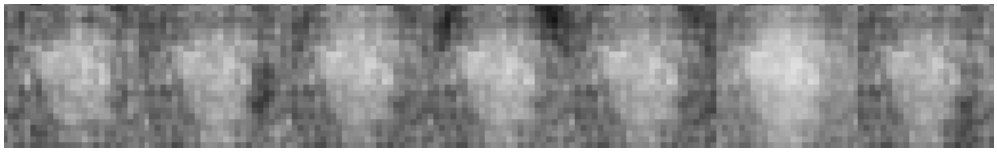
通过设置隐藏层中某一神经元被激活，其余神经元未激活，然后将该层神经元状态逆向传递至可视层，最后将可视层所获得的信息通过灰度图显示出来的方式可以获得该神经元所表示的特征。图 3.3 中（a）、（b）、（c）、（d）四幅子图分别代表 H_1、H_2、H_3、H_4 层中某一神经元被激活时可视层获得的信息。当对 H_i 中的神经元进行激活状态设置时，不考虑 $H_j(j>i)$ 层的作用。



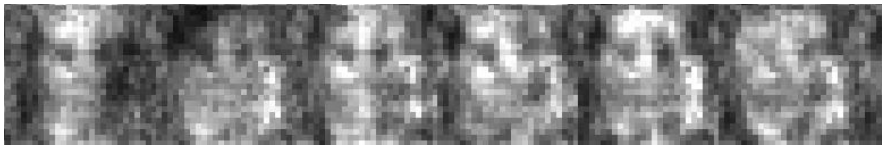
（a）H_1 层示意图



（b）H_2 层示意图



（c）H_3 层示意图



（d）H_4 层示意图

图 3.3 各层神经元被激活时所表示特征示意图

3.3.2 分类层训练

分类层根据 H_4 的输出，获得输入数据的语义信息，实现对输入数据的人脸和非人脸分类。为了使检测网络具有统一性，设定分类层中某一神经元被激活的条件概率分布与 P-RBM 中某一神经元被激活的条件概率分布相同，即

$$P(h_j=1|V)=\frac{1}{1+e^{-(b_j+V^T w_j)}} \tag{3.16}$$

其中 h_j 表示分类层中第 j 个神经元；V 表示 H_4 的输出；b_j 表示分类层中第 j 个神经元的偏置；w_j 表示隐藏层 H_4 与分类层连接权值 W_{class} 的第 j 列。

分类层采用监督训练方式，其流程如图 3.4 所示。

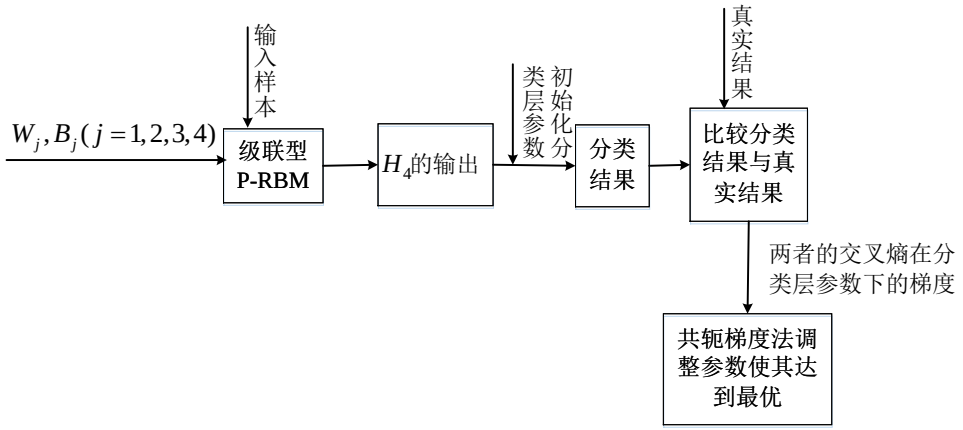


图 3.4 分类层训练流程图

其中 W_j 表示第 $j-1$ 隐藏层与第 j 隐藏层的连接权值；B_j 表示第 j 隐藏层的偏置；分类结果与真实结果之间的交叉熵[44, 61]如下：

$$H(P,\hat{P})=-\sum_i (p_i \log \hat{p}_i+(1-p_i)\log(1-\hat{p}_i)) \tag{3.17}$$

其中 p_i 表示第 i 个数据的真实结果； $\hat{p}_i$ 表示第 i 个数据的分类结果； P 表示所有数据的真实结果的集合； $\hat{P}$ 表示所有数据的分类结果的集合； $\sum(\cdot)$ 表示对所有数据的交叉熵求和。这里采用共轭梯度法[62]调整参数，从而在交叉熵最小时获得最优参数。

共轭梯度法是对梯度下降法[62]的一个修正，由于梯度下降法在接近极小值点处会出现锯齿形搜索路径，使得搜索速度较慢，而共轭梯度法在负梯度方向上叠加一个修正方向来避免锯齿形的搜索路径，加快了搜索速度。其具体过程如下：

1）从初始化参数值 $x^{(1)}=(B,W)$ 开始，计算交叉熵 $H(P,\hat{P})$ 在该点下的梯度 $g_1=((\hat{P}-P),(\hat{P}-P))$。

若 $\|g_1\|\le\sigma$，则 $x^{(1)}$ 为所求的极小值点，即此时的网络参数是最优参数。

若 $\|g_1\|>\sigma$，则在最优搜索步长 λ_1 下沿负梯度方向 $d^{(1)}=-g_1$ 搜索到点 $x^{(2)}=x^{(1)}+\lambda_1 d^{(1)}$。

2）计算交叉熵 $H(P,\hat{P})$ 在点 $x^{(2)}$ 下的梯度 $g_2=((\hat{P}-P),(\hat{P}-P)V)$。

若 $\|g_2\|\le\sigma$，则 $x^{(2)}$ 为所求的极小值点。

若 $\|g_2\|>\sigma$，则在最优搜索步长 λ_2 下沿 $d^{(1)}$ 的共轭方向 $d^{(2)}=-g_2+\beta_1 d^{(1)}$ 搜索到点 $x^{(3)}=x^{(2)}+\lambda_2 d^{(2)}$。

重复步骤 2 直至寻找到极小值点，或者完成所要求的搜索次数。

其中 $\lambda_k=-\dfrac{g_k^T d^{(k)}}{d^{(k)T}d^{(k)}}$ 表示第 k 次搜索时的最优搜索步长； $\beta_k=\dfrac{\|g_{k+1}\|^2}{\|g_k\|^2}$ 表示比例因子； $\|\cdot\|$ 表示取模； σ 表示所要求的精度。

3.3.3 整体优化

由于级联型 P-RBM 训练和分类层训练都是局限于训练单个子网络，因此训练所得的参数只是子网络的最优参数，并不是整个深度学习检测网络的最优参数，这里通过一个整体优化过程对之前训练所得的各个子网络进行整体训练，从而获得整个学习网络的最优参数。之前的级联型 P-RBM 训练构成整体优化的预训练，使得整体优化能有一个较好的初始化参数，解决多隐藏层学习网络整体训练容易陷入局部最优的问题。整体优化过程如图 3.5 所示：

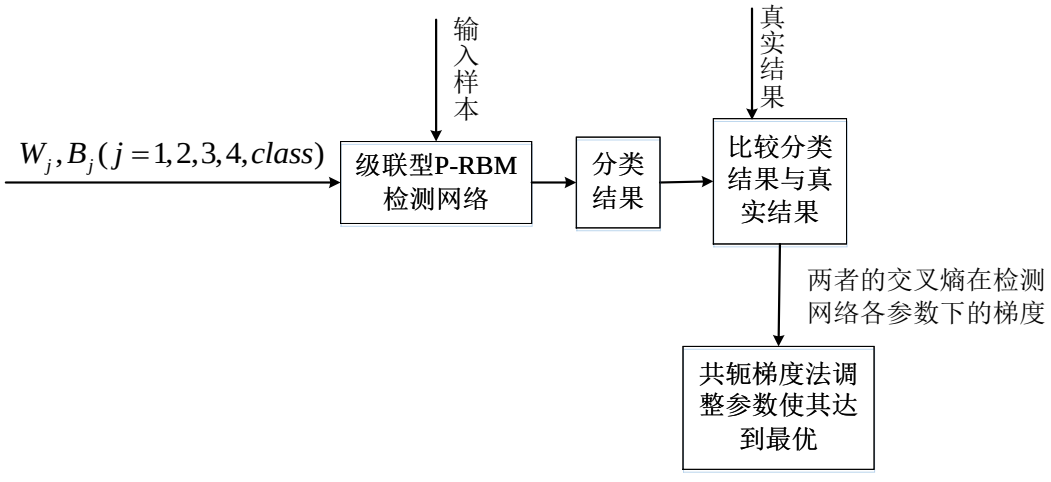
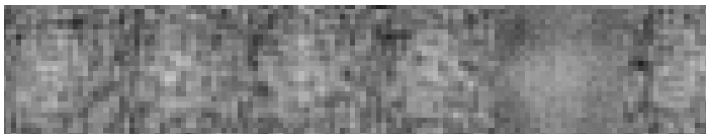


图 3.5 整体优化流程图

整体优化过程与分类层训练十分相似，差别仅在于前面调整的是分类层中的参数，这里调整的是整个检测网络的参数。由于在该检测网络中数据都是从左往右传递，即从可视层向隐藏层传递，因此为了简化训练过程，这里不再对可视层的偏置进行优化。

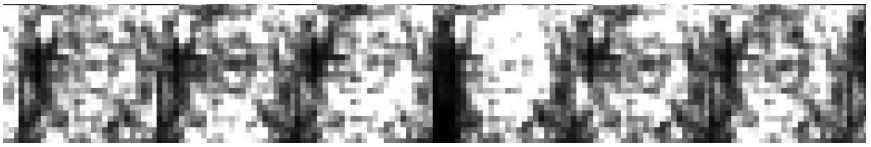
完成整体优化后通过对各隐藏层及分类层中的神经元进行激活状态设置，再利用可视层获得的信息来显示出激活神经元所表示的特征，部分示意图如图 3.6 所示。对比图 3.6 和图 3.3 可以发现，经过整体优化后，各层神经元学习到的人脸特征更加清晰，随着层数地递增，人脸形状越来越明显，特别是对分类层中的神经元进行激活状态设置已经可以获得近乎标准的人脸模板，如图 3.6 中图（e）的左图所示。



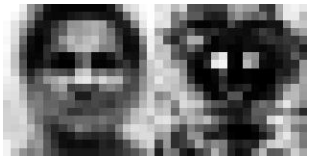
（a）H_1 层示意图

（b）H_2 层示意图

（c）H_3 层示意图



（d）H_4 层示意图



（e）分类层示意图

图 3.6 整体优化后隐藏层及分类层中某一神经元被激活时所表示特征示意图

3.4 基于视频单帧的人脸检测

完成训练的级联型 P-RBM 深度学习检测网络通过对肤色检测所生成的候选区域进行是否为人脸的分类来实现视频单帧的人脸检测。首先读取视频帧图像，并对其进行光照补偿，减少光照对检测效果的影响，同时调整图像尺度，缩小过大的图像使后续检测能够更加快速，接着在 YCbCr 颜色空间上利用肤色椭圆拟合算法[63]进行肤色区域检测，然后将检测得到的候

选区域转换为一维的列向量并逐个输入训练所得的深度学习网络，根据分类结果删去非人脸区域，最后在图像中标定人脸完成人脸检测。

深度学习网络的人脸检测过程如下：

1）对输入数据进行灰度值归一化，并将其赋给可视层V 中各神经元，作为该神经元的激活状态。

2）根据式（3.6）将数据逐层传递至H_4，其中前一隐藏层作为后一隐藏层的输入层。

3）根据式（3.16）获得分类层中各神经元的概率，然后比较分类层中各神经元的概率值，其最大者即为输入数据的分类结果。若分类结果为人脸，则在图中标定出输入数据所对应的区域作为人脸检测的结果；若分类结果为非人脸，则跳过该区域。

3.5 实验与分析

本章随机选取 7000 张来自 LFW 数据集和 4000 张来自 CAS-PEAL 数据集的人脸图，然后截出图像中的人脸区域，构成训练级联型 P-RBM 深度学习检测网络的人脸样本集。其中 LFW 人脸图为自然环境下所拍摄的人脸图，因此部分人脸图像具有一定的旋转角度且各人脸图像之间的面部变化不一；4000 张来自 CAS-PEAL 中的人脸图主要为一些人脸在水平方向或竖直方向旋转角度较大或者是佩戴帽子、眼镜等饰物的人脸图，用于增强学习网络对旋转角度较大或者佩戴饰物的人脸的检测性能。同时，我们制作 10000 张非人脸样本图，作为训练该深度学习网络的负样本。

肤色检测所生成的候选区域需要经过尺度归一化才能作为输入图片输入级联型 P-RBM 深度学习检测网络进行人脸检测，若输入图片太小则深度学习网络无法获得足够的有用信息，若输入图片太大则会带来很多无用信息，给检测造成干扰。因此，本章通过设置候选区域的归一化模板分别为19×19，25×25 以及33×33 来测试输入图片大小对深度学习网络检测性能的影响，从而获得最适合该学习网络的图片尺度，然后在该设置下通过单人脸检测、多人脸检测以及具有旋转角度的人脸检测来测试本章算法的检测性能。

根据图 3.2 的模型结构，对级联型 P-RBM 深度学习检测网络中各层神经元数进行如下设置：可视层V 的神经元数为输入图片（即归一化模板）像素；隐藏层神经元数需要满足两个要求：1）第一隐藏层神经元数大于可视层；2）各隐藏层神经元数逐层递减且各层递减程度大致相同，因此当输入图片为19×19时，学习网络各层神经元数设置为$V=361$，$H_1=450$，$H_2=350$，$H_3=250$，$H_4=100$；当输入图片为25×25时，各层神经元数设置为$V=625$，$H_1=700$，$H_2=450$，$H_3=250$，$H_4=100$；当输入图片为33×33时，各层神经元数设置为$V=1089$，$H_1=1200$，$H_2=650$，$H_3=350$，$H_4=100$；分类层只需进行人脸和非人脸的分类，因此设置为 2 个神经元。

3.5.1 输入图片大小设置

通过对 LFW 数据集和 PKU-SVD-B 数据集[64]中的 EAST 数据集进行人脸检测来测试输入图片大小对算法检测性能的影响。其中 EAST 数据集是由连续的 994 帧，大小为1920×1080的

多人脸视频图像组成，共有 2014 张人脸；LFW 数据集是从 LFW 人脸库中随机抽取除训练样本外的 1884 张单人脸图。

表 3.1 输入图片大小实验

数据集	输入图片大小	正确检测率	漏检率	误检率	检测时间
EAST 数据集	19×19	87.4%	12.6%	1.14%	0.20 s/帧
	25×25	90.5%	9.5%	0.49%	0.25 s/帧
	33×33	88.5%	11.5%	0.64%	0.29 s/帧
LFW 数据集	19×19	98.4%	1.6%	1.3%	0.033 s/帧
	25×25	99.6%	0.4%	1.0%	0.035 s/帧
	33×33	97.8%	2.2%	1.6%	0.047 s/帧

从表 3.1 可以看出，当归一化模板为25×25时，本章算法在这两个数据集上的检测性能最优。对于 EAST 数据集，归一化模板为33×33时的检测性能要优于归一化模板为19×19时的检测性能，而对于 LFW 数据集，归一化模板为19×19时的检测性能要优于33×33时的检测性能。这是因为 EAST 数据集中视频图像的大小为1920×1080，而 LFW 数据集中图像的大小为250×250，这使得两个数据集中单个人脸所含像素信息不同，从而导致相同归一化模板的检测性能不同。

为了证明第一隐藏层神经元数大于可视层有利于提高级联型 P-RBM 深度学习检测网络的分类性能，下述实验将测试在$V=625$，隐藏层神经元数设置分别为：1）$H_1=700$，$H_2=450$，$H_3=250$，$H_4=100$（下文用$H_1=700$来表示）；2）$H_1=550$，$H_2=400$，$H_3=250$，$H_4=100$（下文用$H_1=550$来表示）时算法的检测性能。同时，用 RBM 表示将检测网络中的 P-RBM 换成 RBM 后的检测结果，其各层神经元数的设置与$H_1=700$时相同。

3.5.2 单人脸的算法性能测试

通过 FERET 数据集和 LFW 数据集测试算法对单人脸的检测性能。其中 FERET 数据集由 1400 张单人脸图组成。

表 3.2 单人脸的算法性能测试

数据集	检测算法	正确检测率	漏检率
FERET 数据集	Adaboost[65]	96.93%	3.07%
	RBM	96.22%	3.78%
	H_1=550	99.21%	0.79%
	H_1=700	99.85%	0.15%
LFW 数据集	可变形部件模型算法[22]	99.4%	0.6%
	RBM	96.2%	3.8%
	H_1=550	98.9%	1.1%
	H_1=700	99.6%	0.4%

从表 3.2 可以发现，本章算法优于文献[22]和文献[65]的算法。同时，该算法在 FERET 库上的检测时间为 61.54 秒，在 LFW 库上的检测时间为 87.35 秒，基本实现了实时的无误差单人脸检测。部分实验结果如图 3.7 所示。

图 3.7 H_1=700 时单人脸检测结果示意图

3.5.3 多人脸的算法性能测试

通过 PKU-SVD-B 数据集中的 EAST 数据集进行多人脸的算法性能测试。同时将本文算法与目前比较流行的两种人脸检测算法，基于 Adaboost 的人脸检测算法[65]和结合肤色的 Adaboost 人脸检测算法[66]进行检测性能比较（其中 Adaboost 所需分类器来自 OpenCV 库）。

通过对连续的 994 帧，大小为1920×1080 视频图像共 2014 张人脸进行检测得出表 3.3 的实验结果。

部分检测结果图如图 3.8 所示，其中（a）为基于 Adaboost 的人脸检测算法，（b）为结合肤色的 Adaboost 人脸检测算法，（c）为用 RBM 替换本章算法中的 P-RBM，（d）为本章算法。在（a）、（b）、（c）、（d）各部分检测结果示意图中，左右两列图像中的人脸数不同，用以测试算法对不同人脸数量的检测效果；每列分别为三帧连续的视频图像，图像中人脸的位置、旋转角度发生了变化，用以测试算法对移动人脸的检测效果。

表 3.3 各算法性能比较

算法	正确检测率	漏检率	误检率	检测速度
Adaboost[65]	84.1%	15.9%	34.4%	0.95s/帧
肤色+Adaboost[66]	84.5%	15.5%	15.4%	0.61s/帧
RBM	81.9%	18.1%	2.2%	0.27s/帧
H_1=550	85.1%	14.9%	2.1%	0.24s/帧
H_1=700	90.5%	9.5%	0.5%	0.25s/帧

从表 3.3 可以得出，本章算法对于多人脸图的检测，其误检率为 0.5%，漏检率为 9.5%，检测速度能达到 0.25s/帧；同时从图 3.8 可以看出，对于移动人脸以及具有不同人脸数的视频图像其检测效果优于另外两种算法。因此本章算法无论是在检测准确率还是检测速度上都有更好的表现。

（a）基于 Adaboost 的人脸检测算法

（b）结合肤色的 Adaboost 人脸检测算法

（c）RBM

（d）本章算法

图 3.8、各算法检测结果示意图

3.5.4 旋转人脸的算法性能测试

为了测试算法对旋转人脸的检测性能，从 CAS-PEAL 人脸库中分别选取人脸旋转角度为向左旋转 $30^o \sim 60^o$，向左旋转 $60^o \sim 90^o$，向右旋转 $30^o \sim 60^o$，向右旋转 $60^o \sim 90^o$ 的人脸图各 300 张。由于 CAS-PEAL 人脸库中的人脸图为灰度图像，因此该实验不再进行肤色检测，而是直接将人脸图缩小到 25×25 然后输入级联型 P-RBM 深度学习检测网络，通过分类正确率来完成性能测试。部分检测人脸图如图 3.9 所示。

（a）左旋转 $30^o \sim 60^o$　　（b）左旋转 $60^o \sim 90^o$

（a）右旋转 $30^o \sim 60^o$　　（b）右旋转 $60^o \sim 90^o$

图 3.9 各旋转角度人脸示意图

从表 3.4 可以看出，本章算法对于具有一定旋转角度（旋转角度在 30°~60° 之间）的人脸检测准确率很高，对于旋转角度较大（旋转角度在 60°~90° 之间）的人脸也有较好的检测效果，说明该算法对于旋转人脸的检测具有较强的鲁棒性。同时，各组数据库（包含 300 张人脸）的检测时间都在 20 秒以内。

表 3.4、算法对于具有旋转角度的人脸的检测结果

测试类型		正确检测率	漏检率
左旋转 30°~60°	RBM	85.7%	14.3%
	H_1=550	89.6%	10.4%
	H_1=700	93.1%	6.9%
左旋转 60°~90°	RBM	77.3%	22.7%
	H_1=550	82.7%	17.3%
	H_1=700	85.4%	14.6%
右旋转 30°~60°	RBM	86.3%	13.7%
	H_1=550	90.5%	9.5%
	H_1=700	94.8%	5.2%
右旋转 60°~90°	RBM	75.7%	24.3%
	H_1=550	81.4%	18.6%
	H_1=700	86.3%	13.7%

从表 3.2、表 3.3 以及表 3.4 可以得出：1）当 H_1 层神经元数大于可视层 V 的神经元数时，

级联型 P-RBM 深度学习检测网络的漏检率和误检率都有所降低，表明这种设计思路能够达到提高学习网络检测准确率的目的；2）级联型 P-RBM 深度学习检测网络的检测性能优于级联型 RBM 深度学习检测网络，说明 P-RBM 中神经元的概率态表征优于 RBM 中的二元逻辑状态。

3.6 本章小结

视频人脸检测需要解决在多种非理想条件下的检测准确性、鲁棒性，以及检测速度的问题，本章根据人脑神经元所具有的连续分布激活状态，提出一种面向视频单帧人脸检测的级联型 P-RBM 深度学习检测网络。它利用 P-RBM 中神经元的概率状态表征来模拟人脑神经元所具有的一个从最活跃到最不活跃的连续分布激活状态，使其能更好地模拟人脑对信息的响应状况。同时，根据人脑视觉系统分层处理和学习图像信息的过程，以 P-RBM 为核心，通过级联多个 P-RBM 构建深度学习网络，从边缘特征、轮廓特征、局部特征一直到语义特征，逐层递进，根据最后的语义特征获取输入数据所要表达的信息从而准确地实现非理想条件下的人脸检测。另外，为了满足视频人脸检测对检测速度越来越高的要求，本章通过逐层递减各隐藏层神经元数来约束网络中神经元的规模，提高计算效率，并在预处理层利用肤色检测生成候选区域，在缩小检测范围的同时提高神经网络的鲁棒性。检测网络通过预训练来初始化网络参数，并采用逐层贪婪学习避免了多层网络训练由于误差多层传递使得最后一层获得的误差太小不能很好调整网络参数的情况，在保持学习网络全局优化性能的同时进一步提高了鲁棒性。仿真实验表明，该算法对于单人脸和多人脸图都有较好的检测性能，且对于旋转人脸具有较强的鲁棒性。同时，该算法检测速度快，基本满足实时人脸检测的要求。

第 4 章 多帧间信息融合的视频人脸检测算法

4.1 引言

上一章通过级联型 P-RBM 深度学习检测网络来实现基于单帧的非理想条件下快速准确的视频人脸检测，这是目前常用的一类视频人脸检测方法，即将视频分解为多帧静态图像，然后采用人脸检测方法进行检测，这种基于单帧静态图像的检测所利用的特征或分类信息只来源于一张图片。但是，视频相比静态图像，其最大的特点是各帧图像之间具有连续性，这种连续性能够提供包括位置变化、局部区域像素变化以及视频中目标大小等信息，这些信息能在去除背景干扰，设定阈值等方面提供帮助，从而为降低漏检率和误检率做出贡献。然而，并非所有连续性信息都能为检测带来帮助，若采用一些无用的连续性信息，如在已有肤色检测时又利用帧差法获得运动区域，不仅不能降低误检率和漏检率，还会耗费检测时间。因此，如何在现有的单帧人脸检测结果中融入帧间连续性信息，以及提取哪些连续性信息是本章算法的关键所在。

本章依据视频帧间各连续性信息的特点，通过对上一章实验结果中出现的误检和漏检情况进行分析，选取出能对漏检和误检进行修正的有用信息，并制定信息传递规则将修正信息与检测信息相结合，继而提出一种多帧间信息融合的视频人脸检测算法。该算法对于单帧视频图像的人脸检测仍然采用上一章的算法，然后针对各视频帧图像，首先利用人脸肤色区域长宽比允许范围去除部分误检区域，其中范围的设定采用自适应更新方式从而获得检测视频最适宜的边界条件。接着根据前后相邻的一帧或多帧视频图像的检测结果估计当前帧的检测结果，并与当前帧的真实检测结果进行比较，按照事先制定的对比规则对两者的差异进行判断以修正误检和漏检情况，最后在视频帧中标定人脸完成视频人脸检测。该算法不仅保留级联型 P-RBM 深度学习检测网络在检测准确率和检测速度上的优势，并且通过视频帧间的连续性对检测结果进行了修正，使得检测结果更加准确，提高了算法检测性能，同时对局部被遮挡人脸的检测也有一定的改善。而修正过程无需太多计算步骤，只需比较位置坐标间的差异，因此对检测速度几乎无影响。

4.2 多帧间信息的融合

视频帧间的连续性信息包含位置移动信息、对应区域像素值变化信息以及一些用于阈值设置的形状信息、轮廓线长度信息等。其中位置移动信息、像素值变化信息只能通过相邻帧图像中目标的位置变化或是相对位置像素值之差来获得，而形状信息、轮廓线长度信息等虽然能在单张图像中直接提取，但对于不同场景、不同监控设备所拍摄到的视频，需检测目标在图像中所占的区域大小不同，若用同一阈值进行筛选适应性必定很差。因此，利用视频中连续出现的目标来提取该视频最适合的阈值将能更好地实现检测任务。

然而，仅仅提取上述这些信息还不够，重要的是要将这些信息与单帧视频图像的检测结果相结合，如利用连续出现的目标所提供的边界信息来设定最佳阈值条件以去除当前帧中的误检；利用前后帧中正确的检测结果来估计当前帧应有的检测结果以修正当前帧中的漏检和误检，从而真正将前后帧视频图像信息融合进当前帧的检测结果中。本节首先针对级联型 P-RBM 深度学习检测网络对视频单帧的人脸检测结果选取适宜的连续性信息，再根据所选取信息的特性制定多帧间该信息的融合规则，最后将该规则应用到学习网络的检测结果中，进一步提高视频人脸检测的准确率。

4.2.1 连续性信息的选取

除去背景信息，视频中的目标不会突然出现在视频中的某一位置，也不会从某一位置突然消失，而是有一个逐渐移入逐渐移出的过程。就视频中的人脸而言，它在视频中是一个从边缘出现再从边缘消失的变速移动过程。这里的边缘不一定是指图像的边界，也有可能是所拍摄场景中的门、墙等能将人完全遮挡住的物体。因此在连续的多帧视频图像中，人脸的位置坐标虽然会发生变化，但这变化是在一定范围内且和视频中人的移动速度相关。因此，可以在当前帧的检测结果中融入前后一帧或多帧的检测结果以修正当前帧的检测结果，即利用人脸在前后帧图像中的位置信息及移动方向去估计其在当前帧中的位置，然后将该估计位置与检测位置进行对比以判断当前帧对于该人脸的检测结果是否存在差错。

另外，通过对第 3.5.3 节多人脸检测实验中的漏检和误检情况（结果示意图如图 4.1 所示）进行统计分析可以发现，漏检和误检基本不会重复发生在同一目标或区域上，即漏检的人脸在其前后帧图像中会被正确检测出，而误检的区域在其前后帧图像中并不会被误检。由此可以得出级联型 P-RBM 检测网络所产生的漏检和误检基本都是无规律的出现在单帧图像当中，这进一步证明可以利用人脸的移动规律及前后多帧图像的检测结果估计出当前帧中对应的人脸位置，然后将估计结果与检测结果进行比较，从而对漏检和误检进行修正，删去误检，补上漏检，具体修正方式将在 4.2.2 节中进行介绍。

(a)存在漏检的情况

(b)存在误检的情况

图 4.1 级联型 P-RBM 深度学习检测网络人脸检测结果图

对于少数在多帧图像中连续出现的误检区域（如图 4.2 所示，其中红色框为肤色检测结果，黄色框为人脸检测结果），由于在预处理层是采用肤色检测去除背景区域，这使得误检都出现在肤色或类肤色区域，而这些区域在长宽比上很难满足人脸所具有的长宽比要求，因此可以利用人脸肤色区域的长宽比来去除一些误检情况，进一步提高检测准确率。

图 4.2 误检出现在连续多帧图像中的情况
（红色框为肤色检测结果，黄色框为人脸检测结果）

4.2.2 融合前后帧检测信息的视频人脸检测

根据上一节的分析，本节提出一种融合前后帧人脸肤色区域长宽比信息和人脸位置变化信息的视频人脸检测算法，它在级联型 P-RBM 深度学习检测网络的检测基础上通过制定前后帧检测信息的传递规则来修正漏检和误检，达到提高视频人脸检测性能的目的。具体的信息融合包含两个内容：其一，利用连续出现的人脸获取人脸肤色区域长宽比的允许范围，以去除由于肤色检测所产生的部分误检区域；其二，利用人脸在前后帧的位置信息和移动规律估计人脸在当前帧的位置，以判断当前帧对该人脸的检测结果是否正确，若检测错误则进行修正。考虑到连续出现的误检区域会在利用位置变化信息的修正过程中造成干扰，使得误检进一步扩大，因此这里先利用人脸肤色区域长宽比来删除部分误检情况以减少其对后续检测的影响。

由于不同季节，不同场景下人的着装风格不同，人脸肤色区域所满足的长宽比会随之发生变化（如夏天会包含脖子部分）。因此，这里对长宽比的范围设定分为两步：

1）设定人脸肤色区域长宽比初始范围为：$0.8 \le$ 长/宽 ≤ 1.2 。

2）获取在连续五帧图像中都存在的人脸所对应的肤色区域长宽比，当该长宽比超过所设定的范围时，将该肤色区域的长宽比作为新的边界条件更新人脸肤色区域长宽比的允许范围。该步骤持续执行直到完成整个视频人脸检测。

经过长宽比筛选后，利用前后一帧或多帧视频图像中人脸的位置信息和移动规律对当前帧的误检和漏检情况进行修正。虽然人脸在整个视频中是一个变速移动过程，但由于一帧仅占 1/12 秒，且这里考虑的连续帧数较少，为个位数，因此可以将人脸看成是一个匀速移动过程，从而根据前后帧检测结果及结果之间的移动距离和方向估计出当前帧中人脸所在位置，再根据估计位置与检测位置之间的差异对检测结果进行修正，具体的前后帧位置信息对比融合方式如下。

假设检测图像（即当前帧）为 P_n，图 4.3 中最外层的矩形区域，图像中的人脸区域为正方形区域 F_n，则该人脸的移动范围一定是在以正方形 F_n 的中心 O 为圆心，r 为半径的圆 Q_n 内，

其中r与人脸移动速度有关。也就是说，在当前帧P_n的前后帧图像中，圆Q_n区域内一定存在该人脸。若前后帧的圆Q_n内不存在该人脸，则F_n就是由于误检而产生的人脸区域。

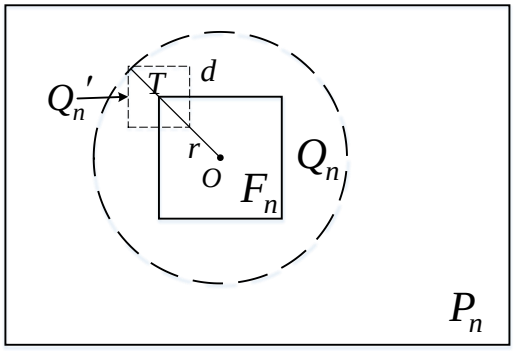


图 4.3 选取区域示意图

为了简化计算复杂度，对于人脸的移动范围不通过圆Q_n来获得，而是通过F_n左顶点T在左右和上下的移动距离获得。假定T的坐标为(x,y)，若前后帧图像在区域$Q_n'(x\pm d/2, y\pm d/2)$（图 4.3 中虚线所画正方形区域为Q_n'，其边长为d）中存在某一人脸区域的左顶点，则认为在当前帧P_n的F_n区域中存在人脸；反之，则认为不存在。由于在整个$(x\pm d/2, y\pm d/2)$区域内逐点搜索比较消耗时间，这里通过前后帧已有的人脸位置信息及相互关系估计出当前帧对应的人脸位置，然后比较该估计位置与当前帧检测位置T之间的差异，并根据对比规则对差异进行判断以修正检测结果。

对比规则及修正方式为：设P_n的前一帧视频图像中某一人脸区域的左顶点T_{-1}的坐标为(x_{-1}, y_{-1})，P_n的后一帧视频图像中某一人脸区域的左顶点T_1的坐标为(x_1, y_1)，

1）当$|(x_1-x_{-1})/2+x_{-1}-x|\le d/2$且$|(y_1-y_{-1})/2+y_{-1}-y|\le d/2$（其中$|\cdot|$表示取绝对值），即当前帧$P_n$的$F_n$区域中存在人脸，而前后帧检测结果显示该区域确实存在人脸时，表明F_n的检测结果正确，则不进行任何修正操作；

2）当$|(x_1-x_{-1})/2+x_{-1}-x|> d/2$或$|(y_1-y_{-1})/2+y_{-1}-y|> d/2$，即当前帧$P_n$的$F_n$区域中存在人脸，而前后帧检测结果显示该区域不存在人脸时，表明在F_n区域上产生了误检，则删除该区域所标定的人脸框；

3）当$|x_1-x_{-1}|\le d$且$|y_1-y_{-1}|\le d$，而区域Q_n'内不存在T，即当前帧P_n的F_n区域中不存在人脸，而前后帧检测结果显示该区域存在人脸时，表明在F_n区域上产生了漏检，则在F_n区域上标定出漏检的人脸。

值得注意的是，比较对象的选取不局限于前后相邻的一帧视频图像，可以选取前后一帧或多帧图像获得一个综合的比较结果。前后帧数选取的越多，准确性越高，但是比较复杂度越大；选取的越少，比较复杂度小但准确性较差。比较间距d的设置需要根据所检测视频进行设定，当视频中的人脸移动速度较快时，比较间距需要适当扩大，否则会造成许多漏检；当视频中的人脸移动缓慢或保持静止时，比较间距需要适当缩小，否则会带来许多误检。同时，比较间距与读取视频帧的间隔数有关，当间隔数较大时，比较间距也要适当扩大。具体设置将在下文实验部分说明。

利用上述对比规则即可对当前帧的检测结果进行修正，删去误检区域，补上漏检区域，最后在视频图像中标定出所有人脸区域完成检测。由于该方法无需太多计算步骤，只需比较各位置坐标间的差异，因此对检测速度影响很小。

4.3 实验与分析

通过 PKU-SVD-B 数据集中的 EAST 视频数据集和实验室自建视频数据集对本章算法进行性能测试。其中 EAST 视频数据集为连续的 994 帧，大小为1920×1080的视频图像，共 2014 张人脸，其人脸较大较清晰，人脸数从 1 到 4 个不等；实验室自建视频数据集为连续的 481 帧，大小为1280×1024的视频图像，共 1920 张人脸，其人脸较小较模糊，人脸数从 1 到 7 个不等。

根据第 3.5.1 节的实验结果，本章设定输入区域的归一化模板为25×25，而级联型 P-RBM 深度学习检测网络仍采用图 3.2 所示的模型结构，因此各层神经元数设置为$V=625$，$H_1=700$，$H_2=450$，$H_3=250$，$H_4=100$；分类层为 2 个神经元。

由于视频人脸检测对实时性有一定的要求，为了在不影响检测速度的前提下获得尽量多的前后帧检测信息，本章选取间隔相同帧数的前后各两帧视频图像的检测结果进行比较，具体比较方式如下表所示：

表 4.1、检测结果比较、修正方式

编号	第一帧	第二帧	第三帧（当前帧）		第四帧	第五帧	修正方式
			估计结果	检测结果			
1	有	有	有	有	有	有	保留
2	有	有	有	有	有	无	保留
3	有	有	有	有	无	有	保留
4	有	有	有	有	无	无	保留
5	有	有	有	无	有	有	补上
6	有	有	有	无	有	无	补上
7	有	有	有	无	无	有	补上
8	有	有	有	无	无	无	补上
9	有	无	有	有	有	有	保留
10	有	无	有	有	有	无	保留
11	有	无	有	有	无	有	保留
12	有	无	无	有	无	无	删去
13	有	无	有	无	有	有	补上
14	有	无	无	无	有	无	保留
15	有	无	无	无	无	有	保留
16	有	无	无	无	无	无	保留

17	无	有	有	有	有	有	保留
18	无	有	有	有	有	无	保留
19	无	有	有	有	无	有	保留
20	无	有	有	有	无	无	保留
21	无	有	有	无	有	有	补上
22	无	有	有	无	有	无	补上
23	无	有	有	无	无	有	补上
24	无	有	无	无	无	无	保留
25	无	无	有	有	有	有	保留
26	无	无	有	有	有	无	保留
27	无	无	有	有	无	有	保留
28	无	无	无	有	无	无	删去
29	无	无	有	无	有	有	补上
30	无	无	无	无	有	无	保留
31	无	无	无	无	无	有	保留
32	无	无	无	无	无	无	保留

“无”表示Q_n'区域内未检测到人脸或估计结果为不存在人脸；“有”表示Q_n'区域内检测到人脸或估计结果为存在人脸；“保留”表示不对当前帧该区域的检测结果进行修正；“删去”表示删去当前帧该区域的人脸框；“补上”表示在当前帧该区域补上人脸框。

4.3.1 不同数据集算法性能测试

通过对 PKU-SVD-B 数据集中的 EAST 数据集和实验室自建视频数据集进行视频人脸检测来测试本章算法的检测性能，视频帧间隔数为 0，即按帧读取视频帧图像，比较间距设置为 13。同时测试仅含位置变化信息和仅含人脸肤色区域长宽比信息时该算法的检测性能，用于测试本章算法的性能优势是来源于其中一者还是两者共同作用。另外，通过与基于 Adaboost 的人脸检测算法[65]、结合肤色的 Adaboost 人脸检测算法[66]以及第 3 章级联型 P-RBM 学习网络人脸检测算法进行检测性能比较来测试本章算法的优异性。（其中 Adaboost 所需分类器来自 OpenCV 库）

从表 4.2 的实验结果可以得出，仅利用长宽比只能删去误检区域，对漏检情况无法进行修正；而仅包含位置变化信息虽然能有效降低漏检率，却会引起误检率的升高，其原因是在制定表 4.1 中的修正方式时更多的考虑漏检的情况，这导致在前后帧同时出现误检时，会将错误信息传递至当前帧。而同时利用长宽比和位置变化信息的方式首先根据长宽比允许范围去除了部分误检，避免了错误信息的传递，在保持较低误检率的同时，利用位置变化信息显著降低了漏检率，提高了算法整体的检测性能。从表 4.2 还可以发现，该算法对检测速度影响很小，基本能满足实时视频人脸检测的要求。

表 4.2 算法性能测试

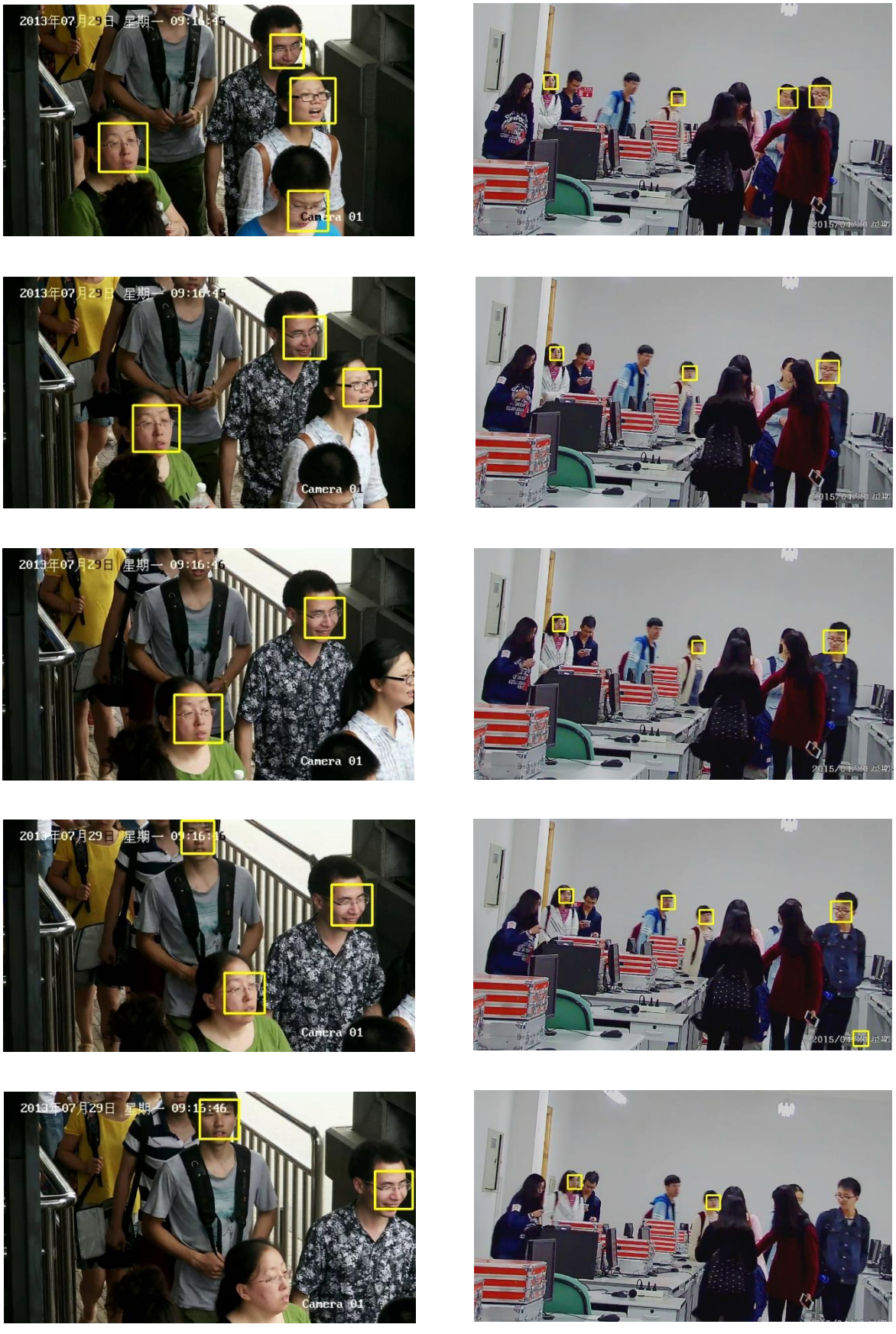
（a）EAST 数据集

算法	EAST 数据集			
	正确检测率	漏检率	误检率	检测速度
Adaboost[65]	84.1%	15.9%	34.4%	0.95s/帧
肤色+Adaboost[66]	84.5%	15.5%	15.4%	0.61s/帧
级联型 P-RBM 学习网络	90.5%	9.5%	0.5%	0.25s/帧
仅含长宽比	89.7%	10.3%	0.35%	0.26s/帧
仅含位置信息	96.0%	4.0%	1.2%	0.27s/帧
长宽比+位置信息	96.6%	3.4%	0.54%	0.27s/帧

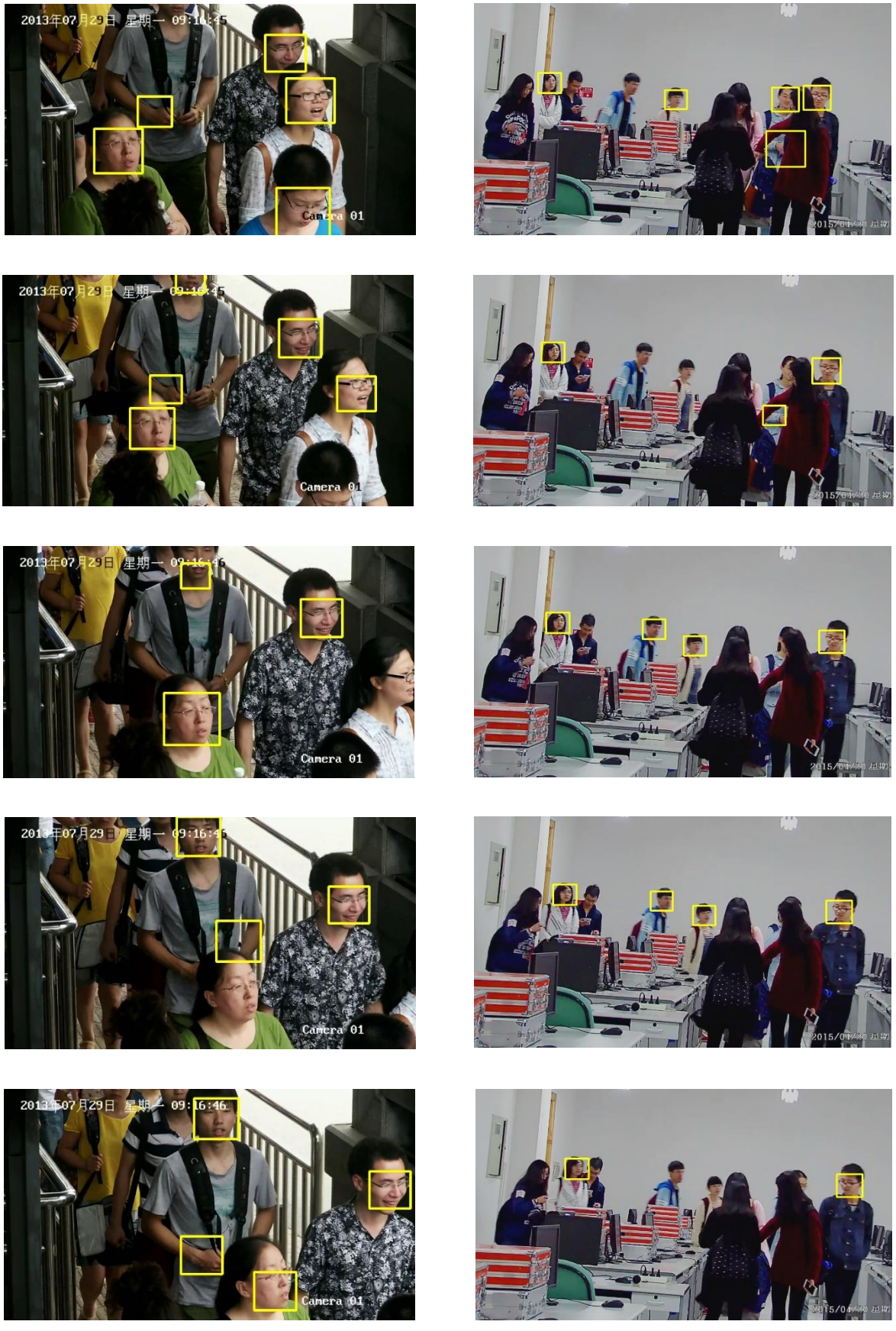
（b）实验室自建视频数据集

算法	实验室自建视频数据集			
	正确检测率	漏检率	误检率	检测速度
Adaboost[65]	80.0%	20.0%	20.9%	2.54s/帧
肤色+Adaboost[66]	83.1%	16.9%	13.8%	0.87s/帧
级联型 P-RBM 学习网络	84.8%	15.2%	6.5%	0.27s/帧
仅含长宽比	84.4%	15.6%	3.5%	0.27s/帧
仅含位置信息	90.4%	9.6%	7.5%	0.28s/帧
长宽比+位置信息	90.8%	9.2%	3.9%	0.28s/帧

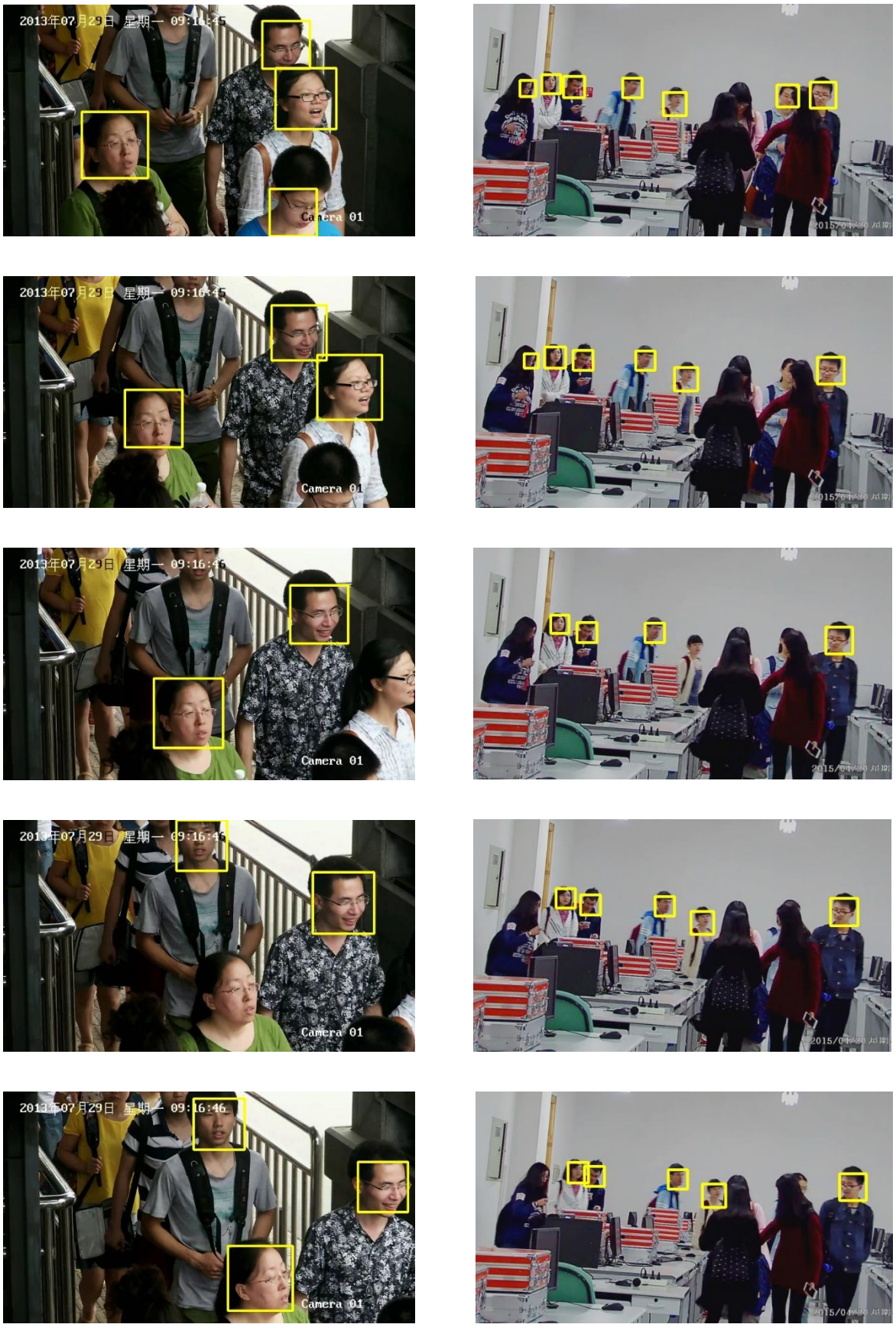
部分实验结果如图 4.4 所示，其中（a）为基于 Adaboost 的人脸检测算法，（b）为结合肤色的 Adaboost 人脸检测算法，（c）为级联型 P-RBM 学习网络人脸检测算法，（d）和（e）为本章算法。在（a）、（b）、（c）、（d）四部分实验结果示意图中，左列为 EAST 数据集，右列为实验室自建视频数据集，用以测试算法对不同人脸数量的检测效果；每列分别为五帧连续的视频图像，图像中人脸的位置、旋转角度发生了变化，用以测试算法对移动人脸的检测效果；（e）为本章算法删除误检的情况，其中左列为级联型 P-RBM 学习网络检测结果，右列为本章算法检测结果。从实验结果图中还可以发现，本章算法能检测出一些部分被遮挡的人脸，如图 4.4（d）中第三幅图像里被数字遮挡住的人脸。



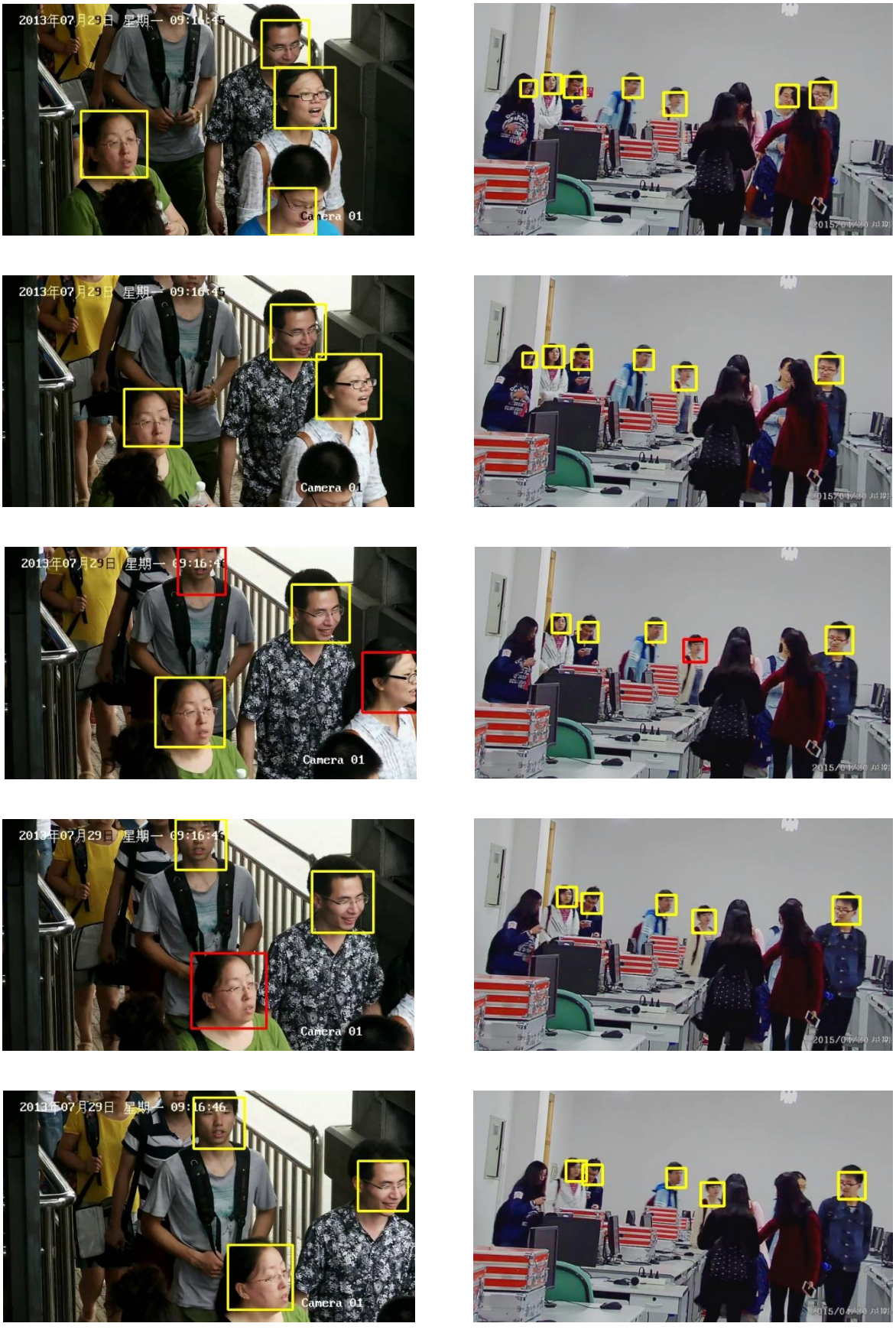
（a）基于 Adaboost 的人脸检测算法



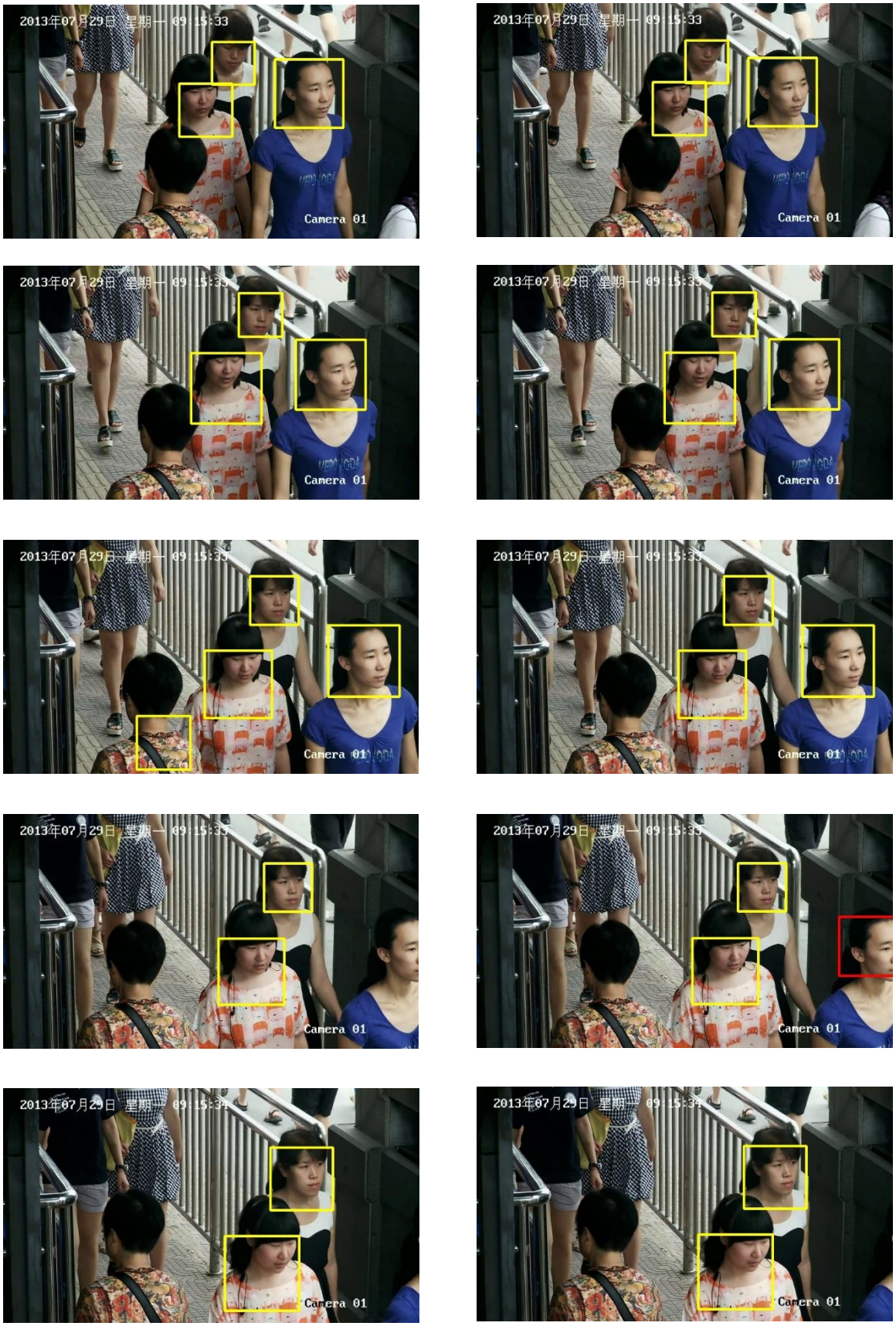
（b）结合肤色的 Adaboost 人脸检测算法



（c）级联型 P-RBM 学习网络人脸检测算法


2013年07月29日 星期一 09:16:45
Camera 01
2013年07月29日 星期一 09:16:45
Camera 01
2013年07月29日 星期一 09:16:45
Camera 01
2013年07月29日 星期一 09:16:45
Camera 01
2013年07月29日 星期一 09:16:46
Camera 01

（d）本章算法（补上漏检）



（左列为级联型 P-RBM 学习网络检测结果，右列为本章算法检测结果）

（e）本章算法（删除误检）

图 4.4、各算法检测结果示意图（红色框为补上的人脸框）

4.3.2 视频帧间隔数和比较间距测试

为了测试比较间距和读取视频帧的间隔数对算法检测性能的影响，本章通过实验室自建视频数据集在不同视频帧间隔数、不同比较间距下的对比实验进行测试，测试结果如表 4.3 和表 4.4 所示。

表 4.3 相同比较间距下的检测性能

间隔数	比较间距	正确检测率	漏检率	误检率
无间隔	13	90.8%	9.2%	3.9%
间隔一帧	13	91.4%	8.6%	3.2%
间隔二帧	13	88.2%	11.8%	6.3%
间隔三帧	13	82.9%	17.1%	8.9%
间隔四帧	13	80.5%	19.5%	12.5%

表 4.4 不同比较间距下的检测性能

间隔数	比较间距	正确检测率	漏检率	误检率
间隔一帧	13	91.7%	8.3%	3.2%
间隔二帧	20	91.6%	8.4%	7.5%
间隔三帧	25	90.3%	9.7%	13.7%
间隔四帧	30	89.8%	10.2%	18.8%

从表 4.3 可得，当比较间距为 13，间隔一帧时算法的检测性能最佳，这是因为在该设置下算法能更好地修正对局部被遮挡人脸的检测，如图 4.5 中第三行图像内最右边被漏检的人脸。在无间隔时，其对应表 4.1 中第 31 种修正方式，因此无法补上该漏检人脸；而在间隔一帧时，其对应第 13 种修正方式，从而能补上该漏检人脸（前后帧图像中的红色框为修正结果，在对当前帧，即第三行图像的检测结果进行修正时不被考虑）。若间隔数继续增加，漏检率和误检率会随之升高。这是因为间隔数变大而比较间距不变时，人脸的移动距离超过了比较距离，从而无法对漏检和误检进行修正。

从表 4.4 可得，当比较间距随着间隔数的增加逐渐增加时，漏检率基本不变，但误检率逐渐增大，这是因为虽然间隔数和比较间距同时扩大，但由于图像中的人脸具有一定的移动速度，相比较的多帧图像所提供的连续性信息不足以判断人脸的移动方向和距离，从而使补上的人脸框所在位置无法对应真实人脸位置，造成误检，如图 4.6 中的红色框。因此，从表 4.3 和表 4.4 可以得出，视频帧间隔数为一帧时本章算法的检测性能最优。

（a）无间隔　　（b）间隔一帧

图 4.5 比较间距为 13 时无间隔与间隔一帧检测结果对比图

图 4.6 间隔数太大时所引起的误检情况

4.4 本章小结

本章提出一种多帧间信息融合的视频人脸检测算法。该算法首先依据上一章所述算法进行视频单帧人脸检测，然后利用视频帧之间的连续性，采用自适应调整方式不断更新获取所检测视频最适合的人脸肤色区域长宽比允许范围，以去除学习网络产生的部分误检，再根据当前帧前后相邻的一帧或多帧视频图像的检测结果所提供的位置信息和相互关系对当前帧的检测结果进行修正，删去误检区域，补上漏检区域。该算法充分利用级联型 P-RBM 深度学习检测网络在仿真人脑感知视觉上的优势，并且针对该学习网络会无规则的在某一帧视频图像中出现漏检或误检的问题，通过视频帧间的连续性对检测结果进行修正，达到提高算法检测性能的目的。同时，该算法无需太多计算步骤，只需比较估计位置和检测位置间的差异，因此对检测速度影响很小。另外，本章还对不同视频帧间隔数、不同比较间距进行了对比实验，从而获取最优的间隔数及其对应的比较间距。实验结果表明，该算法在保持级联型 P-RBM 学习网络较低误检率和较快检测速度的同时，显著降低了漏检率，提高了检测性能，并且对部分被遮挡人脸的检测也有一定的改善。

第 5 章 总结与展望

5.1 研究内容总结

本文主要研究视频中的人脸检测问题，视频人脸检测的主要难点在于如何在复杂的检测背景下将图像中脸部区域的数据信息稳定地映射到语义人脸。另外，检测时间也是一个不得不考虑的因素。现有的视频人脸检测算法无论是采用传统人脸检测技术还是利用运动目标检测技术，通常仅能在一种非理想条件（如复杂背景、光照异常、人脸旋转等）下获得较好的检测效果，当多种非理想条件并存时，检测性能急速下降。但实际的视频检测环境中多种非理想条件并存是常态。针对多种干扰、复杂条件下的视频人脸检测，本文引入深度学习理论并结合视频帧间的连续性信息研究了具有较低漏检率和误检率、较强鲁棒性以及较快检测速度的视频人脸检测方法，为视频人脸检测在实际生活和安防领域的应用提供了理论支持。主要研究成果有以下两个方面：

首先，基于深度学习理论和人脸检测神经网络，提出了一种面向视频单帧的级联型 P-RBM 深度学习人脸检测算法。它首先提出概率态受限玻尔兹曼机，利用其神经元的概率态表征来模拟人脑神经元所具有的连续分布激活状态。然后以此为核心，通过级联多个 P-RBM 构造具有一层可视层，四层隐藏层和一层分类层的级联型 P-RBM 深度学习检测网络，以仿真人脑对视觉的分层次学习模式，实现对输入数据各层次特征的逐层递进提取以及对各层间非线性映射的学习。学习网络中的可视层和四层隐藏层所构成的四个 P-RBM 对应了人脑对图像的四层处理过程，各隐藏层神经元数逐层递减以去除冗余信息，控制网络规模，提高学习网络的检测速度。另外，在预处理层采用肤色检测生成候选区域作为学习网络的输入数据，减少检测范围，进一步提高检测速度。该级联型学习网络通过预训练来初始化网络参数，运用逐层贪婪学习来避免训练误差的多层传递，解决了多层网络训练容易陷入局部最优的问题，再结合整体优化较好的缓解了鲁棒性和准确性之间的矛盾。该算法在充分提取输入数据各层次特征的基础上获得其语义特征，进而实现了较准确的人脸检测任务。

其次，上述这种面向视频单帧的人脸检测算法并未利用视频特有的帧间连续性信息，因此，在上述研究基础之上，提出了多帧间信息融合的视频人脸检测算法。该算法首先分析视频中可以被利用的连续性信息，然后针对级联型 P-RBM 学习网络所产生的漏检和误检情况选取位置变化信息和人脸肤色区域长宽比阈值信息对检测结果进行修正。首先利用人脸肤色区域长宽比允许范围去除部分误检区域，在降低误检率的同时减少误检对后续修正过程的干扰。其中，人脸肤色区域长宽比阈值的设定除了给予初始值外，还利用视频中连续出现的人脸所对应的肤色区域长宽比对其不断更新以获取检测视频最适宜的边界条件。接着，根据前后帧的检测结果以及相互关系估计当前帧对应人脸的位置，然后比较估计位置与检测位置之间的

差异，并将其对应到事先制定的正确检测、漏检情况、误检情况的对比规则中实现对检测结果的修正，进一步提高了视频人脸检测的准确率。该算法不仅充分利用视频帧之间的连续性，并且针对学习网络的不足之处选取有用信息，最重要的是，它通过自适应人脸肤色区域长宽比阈值更新规则和人脸位置差异对比规则很好的将前后帧检测信息融入到当前帧的检测结果当中，达到提高检测性能的目的。

实验数据表明，基于级联型 P-RBM 深度学习网络的人脸检测不仅具有较低误检率和漏检率，而且检测速度较快，同时对于旋转人脸的检测具有较强鲁棒性。将其进一步与多帧间信息融合算法相结合实现的视频人脸检测在保持较低误检率和较快检测速度的同时显著降低了漏检率，还提高了对部分被遮挡人脸的检测性能。

5.2 展望

本文主要针对视频监控下的人脸检测问题，提出了一个在多种非理想条件下具有较好检测性能的视频人脸检测算法，该算法能为智能化设备、智慧安防等应用提供核心的技术支持。虽然针对视频人脸检测问题的研究已经取得了不错的成果，但仍然有许多问题需要进一步解决和完善，下面在总结全文不足之处的基础上，提出后续研究的方向：

1）本文通过级联 P-RBM 来构建深度学习网络，学习网络的输入是一维向量，但是原始图像是二维向量，将其转换成一维向量会丢失一些位置信息，因此在后续的研究过程中应该借鉴卷积理论实现对二维数据的直接特征提取。

2）目前所进行的视频检测都是针对白天拍摄的视频，其光照比较明亮，但是在实际应用中也需要在夜晚的环境下进行检测，若仍利用肤色检测将无法有效地提取候选区域。而学习网络自身没有搜索窗口功能，只能对输入图像进行分类，不能自动获得候选区域。因此如何使得学习网络具备搜索窗口功能而又不会消耗太多检测时间将是另一个需要攻克的难题。

3）视频帧间信息的传递和对比规则都是人为设定的，这会带来一些局限性，可以考虑利用目标跟踪算法对图像中的运动目标建立轨迹模型以更好的预测目标走向。

4）当视频中人脸数较多时，会出现人脸重叠的情况，现有的检测技术只能将重叠人脸看成一个整体进行检测，并不能像人眼一样对其进行分割。因此，重叠人脸的有效检测既是今后人脸检测的一大难点，也是一大研究重点。

致　　谢

在本论文将要完成之际，两年半的研究生生活也接近了尾声。回首两年半的研究生求学生涯，得到了太多人的帮助与鼓励，在此我要向陪伴我、鼓励我及支持我的人致以最诚挚的感谢。

首先我要衷心感谢我的导师叶学义副教授，没有叶老师的耐心指导及谆谆教诲就没有我研究生期间的每一点进步，正是在叶老师的耐心引导下才使得我的研究生课题得以顺利的进行。叶老师严谨的科研态度、渊博的学术知识、诲人不倦的师德品格、和蔼可亲的生活态度不仅使我在研究生学习期间受益良多，也是我今后工作和生活中学习的榜样。生活上叶老师也给予了无微不至的关怀，使我能在学习及生活中克服很多困难，感受到了家一般的温暖。在此特向叶老师表达我最诚挚的感谢和最崇高的敬意。

感谢实验室的张静师姐，高真师姐、张维笑师兄，周天琪师兄，宋倩倩同学，惠舒芸同学等给我的帮助和支持，与他们在一起探讨科研非常的轻松愉快，也是人生中一段难忘的美好回忆。

特别感谢我的室友邹申申，她让我有一个和谐的寝室环境，陪我度过了两年多难忘美好的研究生时光，我们现在已经成为了一生的挚友，这也是我研究生期间收获的一笔重要的人生财富。

最需要特别感谢的是我的父亲、母亲，他们在生活上全心全意的照顾我，无怨无悔的付出得以使我顺利的完成学业，他们是我不断进步的动力。

感谢在百忙之中抽出宝贵时间对我的论文进行评审的专家们，由于作者水平有限，文中难免存在纰漏，恳请各位教授与专家对我的论文予以指正。

参考文献

[1] Yang M H, Kriegman D J, Ahuja N. Detecting faces in images: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(1): 34-58.

[2] Hjelmas E, Low B K. Face detection: A survey[J]. Computer Vision and Image Understanding, 2001, 83(3): 236-274.

[3] 隋静. 基于视频图像的人脸检测方法研究[D]. 西安: 西安电子科技大学, 2011: 1-7.

[4] 高正华. 面向智能视频监控的实时人脸检测算法研究[D]. 杭州: 浙江大学, 2008: 1-8.

[5] Moghaddam B, Jebara T, Pentland A. Bayesian face recognition[J]. Pattern Recognition, 2000, 33(11): 1771-1782.

[6] Schneiderman H, Kanade T. A statistical method for 3D object detection applied to faces and cars[C]. Hilton Head Island, USA: IEEE Conference on Computer Vision and Pattern Recognition, 2000:746-751.

[7] 梁路宏, 艾海舟, 何克忠, 等. 基于多关联模板匹配的人脸检测[J]. 软件学报, 2001, 12(1): 94-102.

[8] 刘明宝, 姚鸿勋, 高文. 彩色图像的实时人脸跟踪方法[J]. 计算机学报, 1998, 21(6): 527-532.

[9] Matsuo H, Sato T, Yokoya N. Vehicle Driver Face Detection in Various Sunlight Environments Using Composed Face Images[C]. Stockholm, Sweden: IEEE Computer Society International Conference on Pattern Recognition (ICPR), 2014: 1687-1691.

[10] Conotter V, Bodnari E, Boato G, et al. Physiologically-based detection of computer generated faces in video[C]. Paris, French: IEEE International Conference on Image Processing (ICIP), 2014: 248-252.

[11] Li H, Lin Z, Brandt J, et al. Efficient Boosted Exemplar-Based Face Detection[C]. Columbus, USA: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 1843-1850.

[12] Phillips P J, Moon H, Rizvi S A, et al. The FERET evaluation methodology for face-recognition algorithms[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(10): 1090-1104.

[13] Huang G B, Ramesh M, Berg T, et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments[R]. Massachusetts, USA: Amherst College, 2007: 09-49.

[14] Gao W, Cao B, Shan S, et al. The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations[J]. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2008, 38(1): 149-161.

[15] Reisfeld D, Yeshurun Y. Robust detection of facial features by generalized symmetry[C]. The Hague, Holland: In Pattern Recognition Computer Vision and Applications, 1992: 117-120.

[16] Yang G, Huang T S. Human face detection in a complex background[J]. Pattern Recognition, 1994, 27(94): 53-63.

[17] 卢春雨，张长水，闻芳，等. 基于区域特征的快速人脸检测法[J]. 清华大学学报:自然科学版，1999, 39(1): 102-105.

[18] 杨秋芬，桂卫华，周书仁. 基于特征三角形的多姿态视频图像人脸跟踪[J]. 计算机工程，2006，32(15): 39-40.

[19] Brunelli R, Poggio T. Face Recognition: Features Versus Templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15(10): 1042-1052.

[20] Leung T K, Burl M C, Perona P. Finding Faces in Cluttered Scenes Using Labeled Random Graph Matching[C]. Cambridge, MA, USA: International Conference on Computer Vision, 1995: 637-644.

[21] Yuille A L, Hallinan P W, Cohen D S. Feature extraction from faces using deformable templates[J]. International Journal of Computer Vision, 1992, 8(2): 99-111.

[22] 尹雪聪. 基于可变形部件模型的人脸检测方法研究[D]. 西安: 西安电子科技大学, 2012: 1-45.

[23] Anvar S M H, Wei-Yun Y, Eam K T. Multiview Face Detection and Registration Requiring Minimal Manual Intervention[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2484-2497.

[24] Ishii H, Fukumi M, Akamatsu N. Face detection based on skin color information in visual scenes by neural networks[C]. Tokyo, Japan: IEEE International Conference on Systems, Man and Cybernetics, 1999: 350-355.

[25] Kjeldsen R, Kender J. Finding skin in color images[C]. Killington, VT, USA: International Conference on Automatic Face and Gesture Recognition, 1996: 312-317.

[26] Jebara T S, Pentland A. Parametrized Structure from Motion for 3D Adaptive Feedback Tracking of Faces[C]. San Juan, Argentina: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997: 144-150.

[27] Propp M, Samal A, Propp M, et al. Artificial Neural Network Architectures for Human Face Detection[J]. Supervised Learning of Fuzzy Artmap Nn's Through Pso, 2006, 1522(1): 196-204.

[28] Rowley H A, Baluja S, Kanade T. Rotation invariant neural network-based face detection[C]. Santa Barbara, CA, USA: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998: 38-44.

[29] Turk M, Pentland A. Eigenfaces for Recognition[J]. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.

[30] Osuna E, Freund R, Girosi F. Training Support Vector Machines: an Application to Face Detection[C]. San Juan, Argentina: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997: 130-136.

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

[32] 黄禹馨，黄山，张洪斌. 基于肤色和器官定位的实时人脸检测[J]. 计算机工程与科学，2014，36(5): 936-940.

[33] 周瑾，王元庆，范科峰. 实时抗干扰的人脸检测方法[J]. 计算机工程与设计, 2013, 34(4): 1399-1403.

[34] Sung K, Poggio T. Example-based learning for view-based human face detection[J]. IEEE Transactions on

Pattern Analysis and Machine Intelligence, 1998, 20(1): 39-51.

[35] Louis W, Plataniotis K N. Frontal face detection for surveillance purposes using dual Local Binary Patterns features[C]. Hong Kong，China: IEEE International Conference on Image Processing (ICIP), 2010: 3809-3812.

[36] 杨辉. 基于Mean Shift算法的运动目标检测与跟踪[D]. 武汉: 武汉工程大学, 2013: 3-18.

[37] 谢仪，鲍可进. 智能视频监控中人脸检测的研究与实现[J]. 计算机测量与控制，2013，21(11): 2921-2923.

[38] 向桂山，王宣银，梁冬泰. 基于人脸肤色和特征的实时检测跟踪算法[J]. 光电工程, 2007, 34(4): 44-48.

[39] Xiong B, Fan X, Zhu C, et al. Face Region Based Conversational Video Coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(7): 917-931.

[40] 龚卫国，桂祖宏，李正浩，等. 融合Adaboost和光流算法的视频人脸实时检测[J]. 仪器仪表学报，2008, 29(7): 1398-1402.

[41] Chang Y C, Lin Y Y, Liao H M. Multi-view face detection in videos with online adaptation[C]. Melbourne, VIC, Australia: IEEE International Conference on Image Processing (ICIP), 2013: 3949-3953.

[42] Lee S T, Mumford D, Romero R, et al. The role of the primary visual cortex in higher level vision[J]. Vision Research, 1998, 38(15): 2429-2454.

[43] 郑胤，陈权崎，章毓晋. 深度学习及其在目标和行为识别中的新进展[J]. 中国图象图形学报，2014, 19(2): 175-184.

[44] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[45] 余凯，贾磊，陈雨强，等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804.

[46] Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex[J]. Journal of the Optical Society of America A: Optics and Image Science and Vision, 2003, 20(7): 1434-1448.

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

[48] 张国翊，胡铮. 改进BP神经网络模型及其稳定性分析[J]. 中南大学学报：自然科学版，2011，42(1): 115-124.

[49] Hinton G E, Dayan P, Frey B J, et al. The "wake-sleep" algorithm for unsupervised neural networks[J]. Science, 1995, 268(5214): 1158-1161.

[50] Luo P, Tian Y, Wang X, et al. Switchable Deep Network for Pedestrian Detection[C]. Columbus, OH, USA: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 899-906.

[51] Nakashika T, Takiguchi T, Ariki Y. Voice Conversion Using RNN Pre-Trained by Recurrent Temporal Restricted Boltzmann Machines[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2015, 23(3): 580-587.

[52] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural

Networks[C]. Lake Tahoe, NV, USA: Annual Conference on Neural Information Processing Systems, 2012: 1097-1105.

[53] 程文博, 张云, 周华民, 等. 基于卷积神经网络的注塑制品短射缺陷识别[J]. 塑料工业, 2015, 43(7): 31-34.

[54] Sun Y, Wang X, Tang X. Deep Learning Face Representation from Predicting 10,000 Classes[C]. Columbus, OH, USA: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 1891-1898.

[55] Smolensky P. Neural and Conceptual Interpretations of Parallel Distributed Processing Models [R]. Cambridge, MA, USA: MIT Press, 1986: 194-281.

[56] 齐敏. 模式识别导论[M]. 北京: 清华大学出版社, 2009: 221-236.

[57] 乔晓艳, 李刚, 董有尔, 等. 弱激光诱导神经元兴奋性改变的实验研究[J]. 物理学报, 2008, 57(2): 1259-1265.

[58] Smith A. A Gibbs sampler on the n-simplex[J]. Annals of Applied Probability an Official Journal of the Institute of Mathematical Statistics, 2014, 24(1): 114-130.

[59] Hinton G E. Training products of experts by minimizing contrastive divergence[J]. Neural Comput, 2002, 14(8): 1771-1800.

[60] Hinton G E. A Practical Guide to Training Restricted Boltzmann Machines[M]. Berlin Heidelberg, Germany: Springer, 2012: 599-619.

[61] Hinton G E. A Practical Guide to Training Restricted Boltzmann Machines[M]. Berlin Heidelberg, Germany: Springer, 2012: 599-619.

[62] 吴春杰. 求解非线性方程组的两类共轭梯度法[D]. 开封: 河南大学, 2014：1-28.

[63] 高建坡, 王煜坚, 杨浩, 等. 一种基于KL变换的椭圆模型肤色检测方法[J]. 电子与信息学报, 2007, 29(7): 1739-1743.

[64] 北京大学视频编码技术国家工程实验室. PKU-SVD-B Database[DB/OL]. [2013]. http://www.smartcity-competition.com.cn/.

[65] 柯丽, 温立平. 改进的AdaBoost人脸检测方法[J]. 光电工程, 2012, 39(1): 113-118.

[66] 杨新权. 基于肤色分割及连续Adaboost算法的人脸检测研究[D]. 成都: 电子科技大学, 2013: 9-57.

附　　录

作者在读期间发表的学术论文及参加的科研项目

发表的论文：

[1] 叶学义, 陈雪婷, 陈华华, 等. 级联型 P-RBM 神经网络的人脸检测[J]. 中国图象图形学报, 2016, 21(7): 875-885.

[2] Ye X, Chen X, Chen H, et al. Deep Learning Network for Face Detection[C]. Hangzhou, Zhejiang, China: IEEE International Conference on Communication Technology, 2015: 504-509.

[3] 张静, 叶学义, 张维笑, 陈雪婷. 一种新的挖掘眼部结构特征的人眼精定位方法[J]. 计算机工程与应用, 2016, 52(12): 158-162.

[4] Ye X, Chen X, Deng M, et al. A multiple-level DCT based robust DWT-SVD watermark method[C]. Kunming, Yunnan, China: IEEE International Conference on Computational Intelligence and Security, 2014: 479-183.

[5] Ye X, Chen X, Deng M, et al. A SIFT-based DWT-SVD blind watermark method against geometrical attacks[C]. Dalian, Shenyang, China: IEEE International Congress on Image and Signal Processing, 2014: 323-329.

发明专利：

[1] 叶学义, 陈雪婷, 陈华华, 顾亚飞, 吕秋云. 基于概率态受限玻尔兹曼机级联的人脸检测方法 (201510788080.6)

[2] 叶学义, 陈雪婷, 邓猛, 汪云路, 何志伟, 赵知劲. 基于 SIFT 的 DWT-SVD 抗几何攻击盲水印方法 (201410146026.7)

参加的科研项目：

[1] 基于视频监控的敏感区域人脸检测系统. 校研究生科研创新基金项目(GK130101299001-081). 2014.3-2015.3.

[2] 面向视频人脸检测的深度学习算法研究. 校研究生优秀学位论文培育基金项目(ZX150605308007). 2015.5-2016.3.