**集美大学计算机工程学院**

**计算机视觉课程论文**

**题 目** 基于MTCNN的实时人脸检测方法及系统

**专 业：**  信息管理与信息系统 **班 级：** 信管1611

**学生姓名：**  刘佳昇 **学 号：**  201621124025

2019年7月12日

### 基于MTCNN的实时人脸检测方法及系统

[摘要] 由于各种姿势，照明和遮挡，在无约束环境中的面部检测和对准具有挑战性。 最近的研究表明，深度学习方法可以在这两项任务上取得令人瞩目的效果。在本文中，我使用了MTCNN深度级联神经网络的多任务框架，该框架由三层网络构成，分别是P-Net、R-Net、以及O-Net。通过分别训练这三个神经网络。将一个网络的输出作为后一个网络的输入。以粗略到精细的方式预测面部的地标位置。指的一提的是，该框架通过选取loss前70%的hard example（比较难训练的样本）执行方向传播，提高了网络的训练效果。经过实验分析，该方法在面部检测中有着卓越精度，同时保持了实时性能。此外，在训练好的模型基础上，设计并开发出一个实时人脸检测系统，适用于门禁，照相以所有需要运用到人脸检测的场景。代码和模型已公开至:

https://github.com/ohMyJason/tensorflow-MTCNN.git

[关键词] 人脸检测； 级联神经网络； 深度学习： MTCNN：

# 目录

[基于MTCNN的实时人脸检测方法及系统 II](#_Toc13794564)

[目录 III](#_Toc13794565)

[第1章 引言 1](#_Toc13794566)

[1.1背景介绍与研究意义 1](#_Toc13794567)

[1.2研究内容与目标 4](#_Toc13794568)

[第2章 深度学习及MTCNN深度级联神经网络原理 5](#_Toc13794569)

[2.1 深度学习简介 5](#_Toc13794570)

[2.1.1 Proposal Network (P-Net)层 5](#_Toc13794571)

[2.1.2 Refine Network (R-Net)层 5](#_Toc13794572)

[2.1.3 Output Network (O-Net)层 6](#_Toc13794573)

[2.2小结 6](#_Toc13794574)

[第3章 人脸检测算法 7](#_Toc13794575)

[3.1 人脸检测算法 7](#_Toc13794576)

[3.1.1 面部分类： 7](#_Toc13794577)

[3.1.2 边界框回归 7](#_Toc13794578)

[3.1.3 面部地标定位 7](#_Toc13794579)

[3.1.4 多源训练 8](#_Toc13794580)

[3.2 小结 8](#_Toc13794581)

[第4章 系统设计与实验结果分析 9](#_Toc13794582)

[4.1系统框架设计 9](#_Toc13794583)

[4.2前端界面设计 9](#_Toc13794584)

[4.3系统实现 9](#_Toc13794585)

[4.3.1 环境配置 9](#_Toc13794586)

[4.3.2 前端界面实现 10](#_Toc13794587)

[4.3.4功能逻辑实现 11](#_Toc13794588)

[4.4实验步骤与分析 12](#_Toc13794589)

[4.4.1 实验设置 12](#_Toc13794590)

[4.4.2 数据集 12](#_Toc13794591)

[4.4.3 评估指标 13](#_Toc13794592)

[4.4.4 实验结果分析 14](#_Toc13794593)

[4.4.4.1多种方法对比 14](#_Toc13794594)

[4.4.4.2多种环境对比 14](#_Toc13794595)

[4.5小结 17](#_Toc13794596)

[结 论 18](#_Toc13794597)

[参考文献 19](#_Toc13794598)

# 第1章 引言

## 1.1背景介绍与研究意义

自动人脸检测技术是所有人脸影像分析衍生应用的基础，这些扩展应用细分有人脸识别、人脸验证、人脸跟踪、人脸属性识别，人脸行为分析、个人相册管理、机器人人机交互、社交平台的应用等。

从应用领域上可以分为：①以企事业单位管理及商业保密为主的商用人脸检测；②大规模联网布控的多角度多背景的安防人脸检测；③反恐安全、调查取证、刑事侦查为主的低分辨率尺度多样的军用/警用人脸检测；④当然还有基于互联网社交娱乐应用等的一般人脸检测。在学术研究中分为约束环境人脸检测和非约束环境人脸检测，如下图。

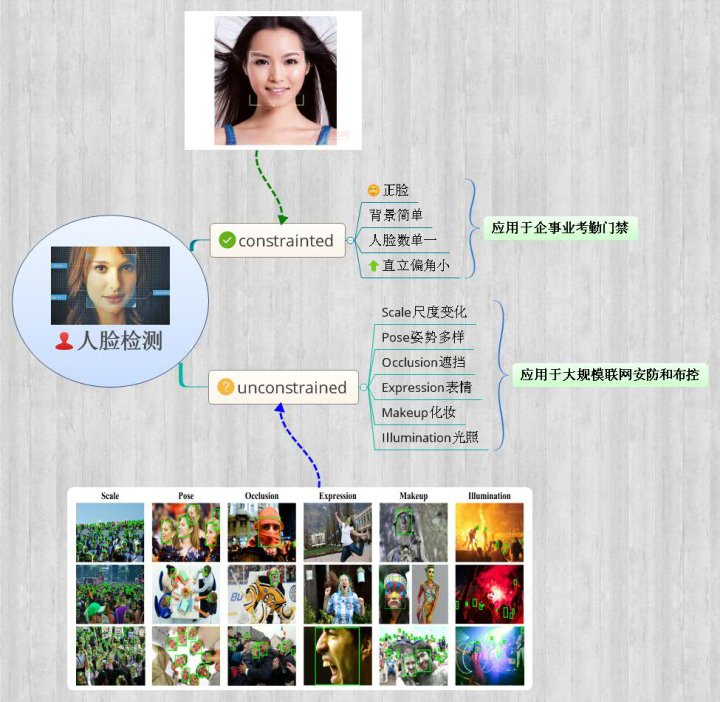


图1-1 度量学习和分类问题区别

在实际中，作为安防企业，人脸检测技术的研发应用在两方面：

1.基本的人脸考勤和门禁等，这一类属于有约束情况，用传统改进算法足以满足性能；

2.安防实时监控，智能视频人脸分析，海量人脸搜索验证、人群数量统计，防踩踏预警等，这类属于非约束情况，不仅对检测算法的精度（包括误检率）要求很高，而且要确保实时性。

调研发现，人脸检测（识别）实战的场景逐渐从室内演变到室外，从单一路况发展到广场、车站、地铁口等。目前基于后者场景，精度若能稳定达到80%以上就属于顶尖技术了（很难达到）。可以想象，雾霾天戴着口罩，冬天戴着帽子，夏天戴着墨镜等，传统算法的检测是比较难的。

例如，公安部门某次招标给出的人脸测试集中一张图如下，可以看出图片模糊、分辨率低，更有戴口罩等大部分遮挡情况。在这种情况下需要识别精度达到一定要求，并且对于召回率（误检率）有一定的保证，传统检测算法是望而却步的，而且目前现有智能视频分析技术的误报率有60%之多，改进的好点的能降低到30%，但还是极大影响实际应用（天安门广场上过识别后因误报高撤了）。综上所述，研究出克服上述困难的人脸检测算法是人脸分析相关应用的首要任务。

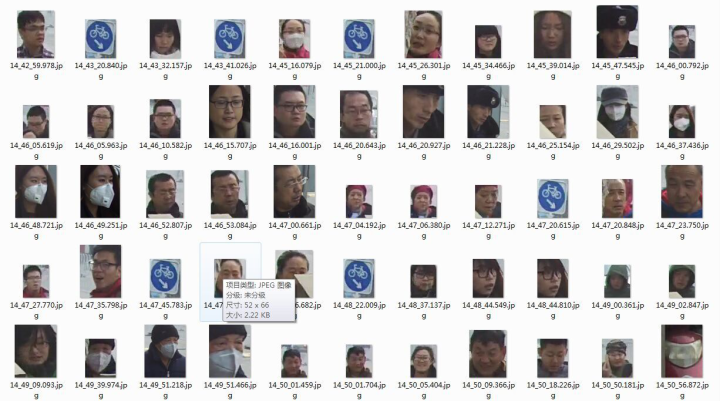


图1-2 公安部提供的测试集

人脸检测的研究在过去二十年例取得了巨大进步，特别是Viola and Jones提出了开创性算法，他们通过Haar-Like特征和AdaBoost去训练级联分类器获得实时效果很好的人脸检测器，然而研究指出当人脸在非约束环境下，该算法检测效果极差。这里说的非约束环境是对比于约束情况下人脸数单一、背景简单、直立正脸等相对理想的条件而言的，随着人脸识别、人脸跟踪等的大规模应用，人脸检测面临的要求越来越高（如上图）：人脸尺度多变、数量冗大、姿势多样包括俯拍人脸、戴帽子口罩等的遮挡、表情夸张、化妆伪装、光照条件恶劣、分辨率低甚至连肉眼都较难区分等。用经典VJ人脸检测器（2010年更新）在非约束评测集FDDB中验证显示：当限定误检数为10个时，准确率不超过10%；为500个时，检测率仅仅为52.8%。所以丞待更好的算法以应用于大规模安防布控等非约束人脸检测场景。

14年底微软美国研究院首席研究员张正友等在CVIU上发表了非约束人脸检测专题综述，文中指出过去十年里，当限定误检数为0或不超过10个时，人脸检测算法的查准率也就是准确率（true positive rate）提高了65%之多（最新基于CNN的算法和传统Vj-boosting算法的对比结果）。

在本文中，我使用了一个新的框架，通过多任务学习使用统一的级联CNN来集成这些任务。该框架的CNN包括三个阶段。

·第一阶段：通过浅CNN快速生成候选窗口。

·第二阶段：通过更复杂的CNN拒绝大量非面部窗口来细化窗口。

·第三阶段：使用更强大的CNN再次细化结果并输出五个面部标志位置。

由于这个多任务学习框架，算法的性能可以显着提高。

本文的主要内容概括如下：

（1）使用了一种新的级联CNN框架，用于联合人脸检测和对齐，并精心设计轻量级CNN架构以实现实时性能。

（2）在具有挑战性的测试中进行了大量实验，与人脸检测中的最新技术相比，显示出使用方法的显着性能和准确性进步。

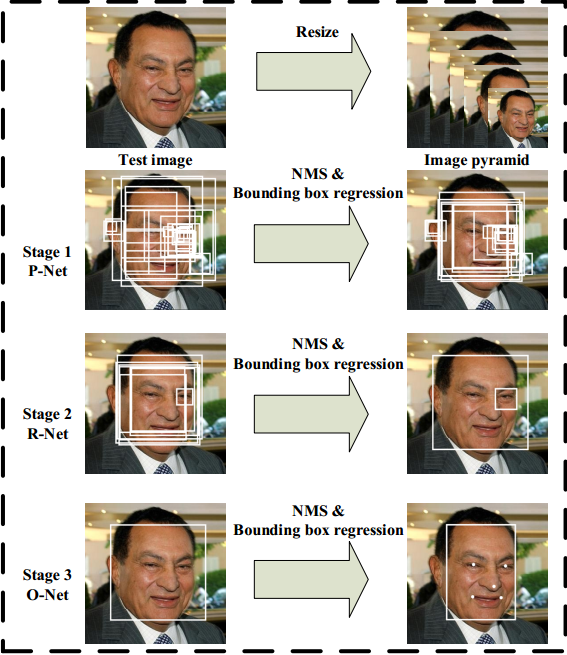


图1-3 MTCNN级联神经网络的人脸检测大致过程

图1-3是级联框架的大致过程，包括三级多任务深度卷积网络。 首先，候选窗口是通过快速提案网络（P-Net）生成的。 之后，通过精炼网络（R-Net）在下一阶段完善这些框。 在第三阶段，输出网络（O-Net）产生最终边界框和面部地标位置。

## 1.2研究内容与目标

本课题研究内容主要是基于级联神经网络优化人脸检测模型，提升人脸检测精度。一是在公开的大规模人脸数据集上训练人脸检测模型；二是采用三层级联神经网络提取特征和对损失函数进行优化。三是基于人脸检测实现的定制人脸检测系统。

本课题目标主要有两个：一方面是，在数据集上基于级联神经网络训练出一个好的人脸检测模型；另一方面是, 在Python上开发构建出一个人脸检测系统，实现对图片或视频中人脸进行实时人脸检测，并显示人脸置信度。

# 第2章 深度学习及MTCNN深度级联神经网络原理

## 2.1 深度学习简介

深度学习(Deep Learning)是机器学习的分支，是一种以人工神经网络为架构，对数据进行特征学习的算法。目前，深度学习结构已有数种，经典代表有深度神经网络、卷积神经网络、递归神经网络等。深度目前已经广泛应用于计算机视觉、自然语言处理、语音识别等领域，并取得了巨大的成果。2016年3月，AlphaGo的表现，展现了深度学习在围棋领域的重大突破。

### 2.1.1 Proposal Network (P-Net)层

该网络结构主要获得了人脸区域的候选窗口和边界框的回归向量。并用该边界框做回归，对候选窗口进行校准，然后通过非极大值抑制（NMS）来合并高度重叠的候选框。

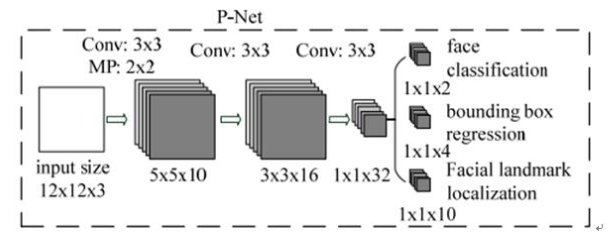


图2-1 Proposal Network (P-Net)层训练过程。其中“MP”表示最大池化，“Conv”表示卷积。 卷积和池化中的步长分别为1和2

### 2.1.2 Refine Network (R-Net)层

该网络结构还是通过边界框回归和NMS来去掉那些false-positive区域。只是由于该网络结构和P-Net网络结构有差异，多了一个全连接层，所以会取得更好的抑制false-positive的作用。

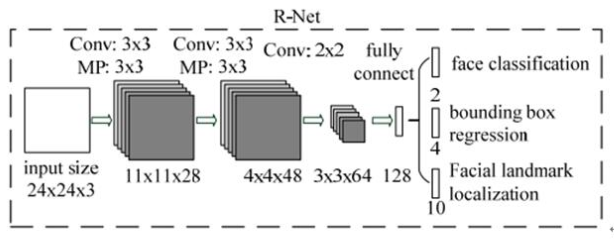


图2-2 Refine Network (R-Net)层训练过程

### 2.1.3 Output Network (O-Net)层

该层比R-Net层又多了一层卷基层，所以处理的结果会更加精细。作用和R-Net层作用一样。但是该层对人脸区域进行了更多的监督，同时还会输出5个地标（landmark）。

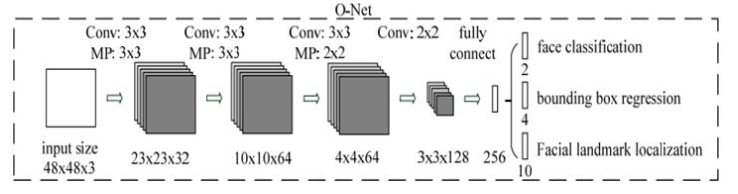


图2-3 Output Network (O-Net)层训练过程

## 2.2小结

通过上文对深度学习基础知识和MTCNN深度级联网络的深入理解和研究。可以看出深度学习在计算机视觉领域有突出表现。MTCNN深度级联神经网络的出现，将多个模型结合在一起，提升了模型的表现能力，有效解决人脸检测中的性能问题。接下来的章节将详细介绍基于深度级联神经网络的人脸检测算法。

# 第3章 人脸检测算法

## 3.1 人脸检测算法

通过以下三项任务来训练MTCNN级联神经网络：

1. 面部/非面部分类
2. 边界框回归
3. 面部地标定位。

### 3.1.1 面部分类：

学习目标被制定为两类分类问题。 对于每个样本，我们使用交叉熵损失：

(1)

其中是网络产生的概率，表明样本是人脸图像的概率。 符号 表示ground-truth(正确的数据标注，其实就是数据集中的人脸位置以及特征位置) 标签

### 3.1.2 边界框回归

对于每个候选窗口，我们预测它与最近的ground-truth 之间的偏移（即边界框的左边，顶部，高度和宽度）。 学习目标被制定为回归问题，我们对每个样本使用欧几里德损失：

(2)

其中是从网络获得的回归目标， 是ground-truth坐标。 有四个坐标，包括左上角，高度和宽度，因此

### 3.1.3 面部地标定位

类似于边界框回归任务，facial landmark（面部地标）检测被公式化为回归问题，我们最小化欧几里德损失：

(3)

其中 是从网络获得的面部地标坐标，是第个样本的地面实况坐标。有五个面部标志，包括左眼，右眼，鼻子，左嘴角和右嘴角，因此

### 3.1.4 多源训练

由于我们在每个CNN中执行不同的任务，因此在学习过程中存在不同类型的训练图像，例如面部，非面部和部分对齐的面部。 在这种情况下，不使用一般损失函数。

例如，对于背景区域的样本，我们仅计算，而另外两个损失设置为0.这可以直接使用样本类型指示符来实现。 然后整体学习目标可以表述为：

(4)

其中N是训练样本的数量，表示任务重要性。

我们在P-Net和R-Net中使用

(5)

而在O-Net中使用

以获得更准确的面部地标本地化 是样本类型指示器。 在这种情况下，采用随机梯度下降来训练这些CNN是很合适的

## 3.2 小结

通过对人脸检测算法流程的研究理解，可以看出MTCNN深度级联神经网络在人脸检测上有重要作用。通过三层CNN训练人脸识别模型，可以让模型提取到更好的特征，可以有效提取出人脸所在的地标。下面进入算法实验分析和系统设计过程中，直观的体会到MTCNN的强大表现。

# 第4章 系统设计与实验结果分析

## 4.1系统框架设计

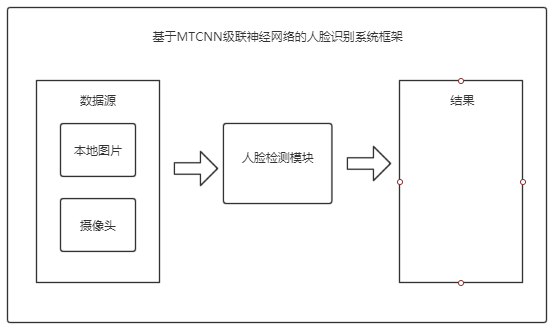


图4-1 人脸检测系统框架

系统框架决定系统开发、维护和扩展的效率。图4-1描述了整个人脸检测系统的框架设计，包括数据源、人脸检测、结果呈现三个模块，用户通过GUI界面进行交互。整个系统首先通过数据源获取图片信息，数据源来自摄像头、本地照片。人脸检测模块是本系统核心模块，将预训练的模型存储在系统中，然后通过提取图片或者视频帧中的人脸特征，将结果呈现在GUI界面中。

## 4.2前端界面设计

系统的GUI设计采用Tkinter实现，主要分为两部分，一部分是图片检测模式，通过上传本地图片完成检测。另外一部分是摄像头检测模式，支持本地摄像头采集人脸信息并实时检测。具体设计如图4-2所示。

## 4.3系统实现

### 4.3.1 环境配置

系统开发环境和算法开发使用环境不一样，因为训练需要大量算力，而部署对配置要求较低。系统开发环境如表4-1所示。

表4-1开发环境

|  |  |
| --- | --- |
| 名称 | 型号 |
| 系统 | Windows10 |
| CPU | Intel I5-8625 |
| GPU | NVIDIA Mx150,2G |
| Tensorflow | 1.8.0 |
| OpenCV | 3.2.0 |
| Tkinter | 3.6.0 |
| PyCharm | 2018 |

### 4.3.2 前端界面实现

界面由图片识别模式和摄像头实时识别模式两个部分构成，方便用户操作。如图4-2所示。界面左上侧部分是显示数据源的图片，右上部分是显示人脸检测结果。下班部分用于显示摄像头实时检测情况，可以手动选择开启或者关闭摄像头显示，方便用户操作。

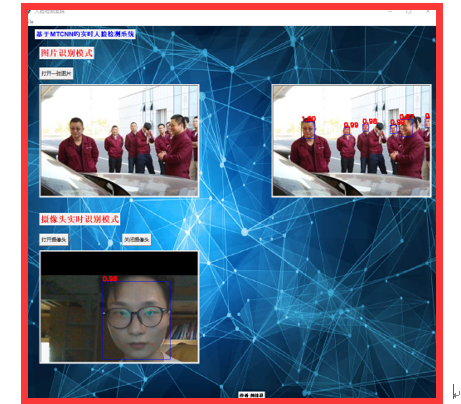


图4-2 人脸识别系统主界面

### 4.3.4功能逻辑实现

功能逻辑是属于中间部件，需要和前端界面交互，又要调用算法进行处理，属于架构里面不可或缺的部分。首先数据源通过OpenCV进行读取，两种不同数据源，其中图片需要用户先选择，文件选择通过TKinter实现。如图4-3所示，图片支持\*.JPG、\*.PNG、\*.GIF、\*.BMP格式。摄像头支持USB、笔记本自带摄像头。



图4-3 两种种选择方式效果图

4.4实验步骤与分析

### 4.4.1 实验设置

本课题算法部分的环境配置如表4-2所示。

表4-2 软件和硬件配置

|  |  |
| --- | --- |
| 名称 | 型号 |
| 系统 | Windows10 64位 |
| CPU | Intel I5-8625 |
| GPU | NVIDIA MX150,2G |
| 内存 | 8G |
| TensorFlow | 1.8.0 |
| OpenCV | 3.2.0 |
| Python | 3.6.5 |

人脸检测训练数据集为WIDER FACE、LFW以及网络爬取的7000张图像。batch size为384，display为100，迭代次数设置为END\_ EPOCH为[30,22,22],学习率减小的迭代次数为LR\_EPOCH=[6,14,20], P-Net对图片缩小倍数stride为2。学习速率设置为0.001，利用TensorFlow在英伟达 MX150显卡上训练人脸检测模型。

### 4.4.2 数据集

本课题采用的数据集有两个分别是WIDER FACE以及[Deep Convolutional Network Cascade for Facial Point Detection的训练集](http://mmlab.ie.cuhk.edu.hk/archive/CNN_FacePoint.htm) 带有的数据集。

WIDER FACE 数据集是香港中文大学开源大规模的人脸检测基准数据集。它包含32203张图片，其中有393703个人脸位置标签，具备有不同尺度，姿势，遮挡，表情，服饰等条件下的图片。如图4-4。



图4-4 WIDER FACE数据集多种环境下的人脸

[Deep Convolutional Network Cascade for Facial Point Detection的训练集](http://mmlab.ie.cuhk.edu.hk/archive/CNN_FacePoint.htm)包含5,590个LFW图像和7,876个从网上下载的图像。训练集和验证集分别在trainImageList.txt和testImageList.txt中定义。这些文本文件的每一行都以图像名称开头，然后是我们的面部检测器修复的面部边界框的边界位置，然后是五个面部点的位置。

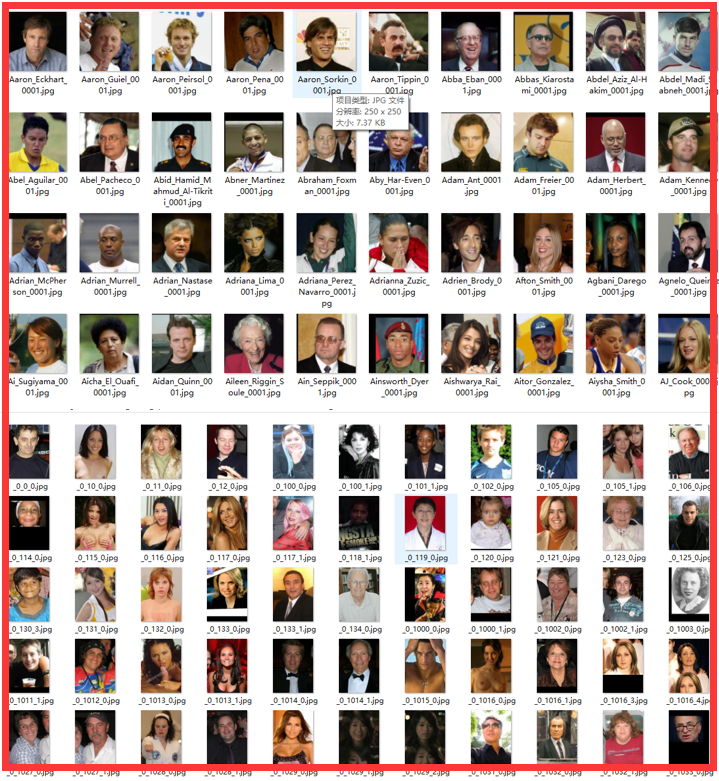


图4-5 LFW数据集以及网络自动获取的多种环境下的人脸

### 4.4.3 评估指标

在人脸检测和人脸识别方面，常用的评估指标有mAP、Precision、Recall、ROC、FAR、F1 score、Top-1、IOU等。根据不同需求选对应指标。本课题采用IOU为人脸检测算法的评估指标。

### 4.4.4 实验结果分析

#### 4.4.4.1多种方法对比

将训练好的模型和多种主流人脸检测方法进行对比，如表4-7所示，人脸检测模型MTCNN在WIDER FACE训练集上表现优异，通过三层训练网络的特征融合、损失函数合理设计，模型效果大大提升，速度达到99帧。

表 4-3 多种方法在WIDER FACE VAL上评估对比情况

|  |  |  |
| --- | --- | --- |
| Method | GPU | Speed |
| Cascade CNN | Titan | 100 |
| Faceness | Titan | 20 |
| DP2MFD | Titan | 0.285 |
| **ours** | **Titan** | **99** |

#### 4.4.4.2多种环境对比

训练好的模型在多种较为严苛的条件下如外物遮挡，水平倾角变换，垂直上倾角变换，垂直下倾角变换，表情变换，环境中增加烟雾等情况下的的人脸检测结果，记录如下：



图4-6 脸部遮挡情况

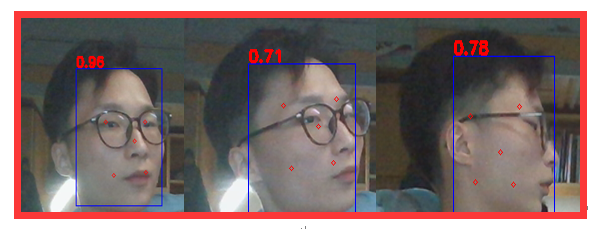


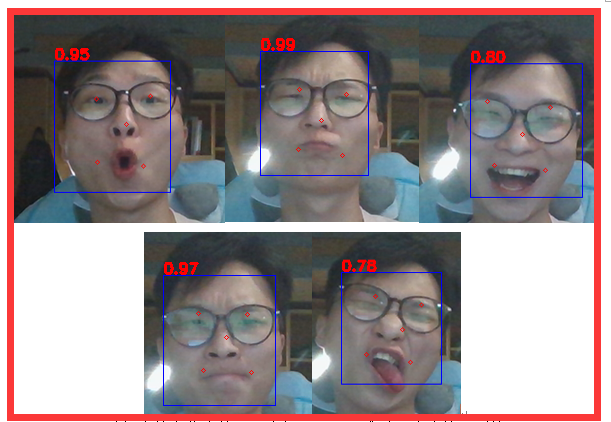
图4-7脸部水平角度变换分别为45度、60度、90度的识别情况



图4-8 脸部垂直水平上倾角分别为45度、60度、70度的识别情况



图4-9 脸部垂直水平下倾角分别为45度，60度的识别情况

图4-10 脸部表情变化为惊讶，生气，开心，难过，疼痛的识别情况

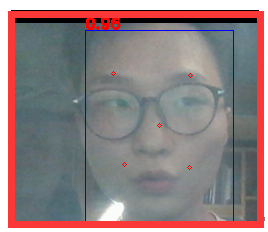


图4-11 识别环境中存在烟雾的识别情况

实验结果显示，在有遮挡物的情况下，左右遮挡可以正常识别，上下遮挡（如遮挡眼睛）无法正常识别。在水平倾角变换分别为45度，60度，90度的情况下均可以正常识别但是特征点显示标定错误。在垂直上倾角分别为45度，60度的情况下识别正常，上倾角提高到70度时，无法识别。在垂直下倾角为45度的时候可以正常死别，但是角度增加至60度时无法正常识别。在脸部存在各种表情的时候可以正常识别。在环境中有烟雾的情况下可以正常识别。

有上述实验结果可以得出结论，系统在检测人脸的过程中具备水平变换，表情变化，环境噪声的鲁棒性，但是在识别上下角度变化以及特征点被遮挡过多的情况下，识别率会收到影响。

## 4.5小结

根据上述实验结果，可以看出MTCNN人脸检测算法相比于其他算法，有着高效，速度快，计算量较小的各方面优势。人脸检测模型识别精度极高。但是也有需要克服的问题：过小的人脸识别精度低、对局部特征识别效果低下。

由上述人脸识别系统可以看出系统将算法和界面融合为一体，可以实现实时的摄像头人脸检测。在多人，遮挡，低光，模糊等条件下有着比较高的匹配精度。

# 结 论

本文从对严苛环境下的人脸检测的需求出发，对基于MTCNN级联神经网络的人脸检测算法进行研究，训练出在各个环境下鲁棒性较高的的人脸检测模型，最后基于Tkinter将算法、界面集成在一起，开发出一款可靠的人脸检测系统。

在人脸检测算法上，采用深度级联神经网络对样本进行分别训练，将上一个网络的输出作为下一个网络的输入，实现高效且准确的人脸检测效果。实验表明这种方法有利于检测不同尺度，不同角度，存在环境噪声的人脸，表现优异。人脸的关键特征检测效果符合人脸识别的标准，实验表明这种结构适用于人脸检测任务，可以有效提取人脸特征，在实时检测中效果好，可用于多种场景下人脸检测。

最后，基于训练好的模型，针对本文需求，开发出了一个人脸检测系统，基本满足各种环境规模下的人脸检测的需求，并且可以用于其他任务。

# 参考文献

1. Deng, L.; Yu, D.Deep Learning: Methods and Applications . Foundations and Trends in Signal Processing[J],2014, **7**: 3–4.
2. Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li: Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks[J],2018,
3. P. Viola and M. J. Jones, “Robust real-time face detection. International journal of computer vision,” vol. 57, no. 2, pp. 137-154, 2004
4. M. T. Pham, Y. Gao, V. D. D. Hoang, and T. J. Cham, “Fast polygonal integration and its application in extending haar-like features to improve object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 942-949.
5. Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, “Fast human detection using a cascade of histograms of oriented gradients,” in IEEE Computer Conference on Computer Vision and Pattern Recognition, 2006, pp. 1491-1498.
6. M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection without bells and whistles,” in European Conference on Computer Vision, 2014, pp. 720-735.
7. J. Yan, Z. Lei, L. Wen, and S. Li, “The fastest deformable part model for object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2497-2504.