文章编号：2095-2163(2021)07-0001-06 中图分类号：TP391.41 文献标志码：A

基于视频的人脸识别校园安全接送系统的研究

姚 砺，周同辉，马 睿，万 燕
（东华大学 计算机科学与技术学院，上海 201620）

摘 要：校园安全接送作为一个严肃的社会问题，需要及时发现不明身份人员混入校园等隐患，而市场上大多数的校园人脸识别系统，基本上都需要专用的硬件设备，且只支持单人逐个检测，这无疑提高了推广的成本。针对以上问题，本文提出了使用普通摄像头的支持多人的人脸识别系统。首先采用 YOLOv3 算法训练人脸数据集，利用 K-Means++算法改进先验框中心位置的预测，提高边界框的准确性，得到人脸检测器，对视频中的行人进行人脸检测；之后利用本文的人脸图像质量评价 FIQUE 算法对人脸进行筛选，增加高质量的人脸图像占比；最后使用 Inception-ResNet-v1 模型提取人脸特征，进行识别。本文方法利用学校现有的摄像头设备，提高了系统的普及率。实验证明了本文多人识别系统的实时性和鲁棒性良好。

关键词：YOLOv3；K-Means++；人脸检测；人脸图像质量评价算法；Inception-ResNet-v1；人脸识别

Face recognition research on campus security transportation system based on video

YAO Li, ZHOU Tonghui, MA Rui, WAN Yan
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 As a serious social problem, campus safe transportation needs to be discovered in time for hidden dangers such as unidentified people entering the campus. Most of the campus face recognition systems on the market basically require dedicated hardware equipment, and only support single-person detection. This undoubtedly increases the cost of promotion. In response to the above problems, this paper proposes a face recognition system that supports multiple people using ordinary cameras. First, use the YOLOv3 algorithm to train the face data set, use the K-Means++ algorithm to improve the prediction of the center position of the prior box, improve the accuracy of the bounding box, and obtain a face detector to detect the faces of pedestrians in the video; then use the FIQUE algorithm for facial image quality evaluation to filter faces and increase the proportion of high-quality face images; finally, use the Inception-ResNet-v1 model to extract facial features for recognition. The method in this paper uses the existing camera equipment of the school to increase the penetration rate of the system. The experiment proves that the multi-person recognition system in this paper has good real-time and robustness.

【Key words】 YOLOv3; K-Means++; face detection; face image quality evaluation algorithm; Inception-ResNet-v1; face recognition

0 引 言

一直以来，校园安全都是一个吸引各界关注的社会热点问题，而中小学上下学的接送问题就是校园安全中的重要一环，并与每位师生、家长都有着密切的关系。2017 年 12 月，“校园安全”入选 2017 年民生热词榜[1]。为此，国务院教育委紧急下发通知，要求各地学校和相关部门加强校园安全管理。同时，公安部也提出准确把握校园安全管理的规律性，建立防控体系，构建长效机制，积极推进平安校园建设[2]。

为方便学生家长和学校教师及时掌握学生的出校情况，本文开发一套基于视频的人脸识别校园安全接送系统，能够准确进行校园接送人员身份识别。相较于传统的核验技术，本文的人脸识别技术[3]能够在无需接触出行者的前提下使用普通摄像头逐帧拍摄人脸照片，结合人脸识别算法，通过人脸数据库进行身份比对，从而对进出校门的人员进行精准识别。

本文是在经典的 YOLOv3[4]目标检测算法的基础上，结合深度神经网络 DNN[5]模块，针对原始 YOLOv3 的 K-Means[6]算法对初始聚类中心的选择不同所产生的聚类结果偏差大的不足，本文采用 K-Means++[7]聚类算法改进先验框中心位置的预测，利用改进之后的算法训练人脸检测模型。为了提高人脸识别的准确率，本文主要采用了下列技术：人脸检测定位后，利用图像质量评价 FIQUE 算法，计算人脸图像质量的评估数值，挑选质量好的人脸照片

作者简介：姚 砺(1967-)，男，博士，副教授，主要研究方向：图像处理；周同辉(1994-)，男，硕士研究生，主要研究方向：图像处理。
通讯作者：姚 砺 Email：yaoli@dhu.edu.cn

收稿日期：2021-05-17

送入预训练的 inception-resnet-v1 模型进行特征提取,该模型是将 Inception[8] 和 ResNet[9] 两者融合。该方法缩短了人脸识别核验的平均耗时,同时提升了人脸识别的感受成功率,人脸识别准确率超过 97%。

1 基于视频的人脸识别方法

基于视频的人脸识别技术应用于校园安全接送,存在的主要问题是视频帧中出现的人脸对象较多,而且受到外界天气、光线以及移动中的行人脸部姿态变化等影响,这都会给人脸识别带来挑战。

本文的研究突破传统的基于专用硬件设备的人脸识别方法,采用普通摄像头拍摄的图像质量往往不如专用的设备,因此本文基于视频的技术对人脸检测的要求较高。图 1 为本文人脸识别的流程。

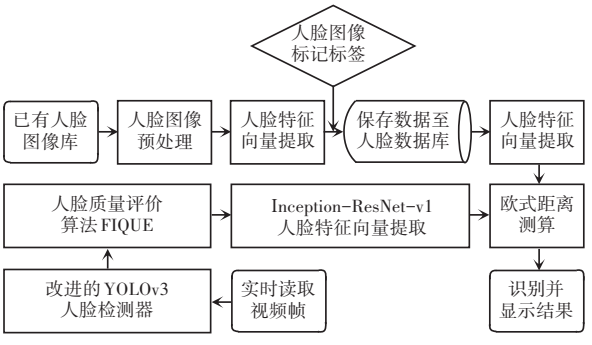


图 1 本文人脸识别流程图

Fig. 1 Face recognition flow chart in this article

为了提高低质量的人脸图像识别的准确性,本文研究了基于 YOLOv3 人脸检测聚类先验框的预测方法,提出使用 K-Means++算法替换原始的 K-Means 进行改进,很好地解决了初始聚类中心选择的敏感性问题。K-Means++算法的基本思想是使初始聚类点中心之间的距离尽可能远,并在每一个维度的输出都是使用 3 个先验框进行预测,因此总共得到 9 种尺度的先验框。同时,基于视频的人脸检测,属于运动物体的检测,存在的问题是对于每一帧获得的一个或多个人脸,无法预知其人脸姿势、人脸质量,因此本文基于 OpenCV 模块提出了人脸图像质量评估算法:利用支持向量和 LIBSVM[10-11] 预测质量得分,从多帧图像中选取符合门限阈值的人脸图像,提升了检测效率和人脸识别的鲁棒性。

基于上述研究的技术路线,本文提出了一套低成本的易于中小学校园推广的基于视频的较高准确率的人脸识别系统,该系统不需要专门的人脸识别硬件设备,可减轻学校的经济压力。

2 基于视频的人脸识别关键技术

2.1 人脸检测算法

如何判断一幅图像或者一帧视频中存在人脸?首次成功地将深度学习应用到目标检测领域的算法是 R-CNN 算法[12],其结构有 2 级网络:首先通过外部区域选择性搜索算法提出可能包含对象的候选边界框,然后将这些区域传递到 CNN[13] 进行分类。Fast R-CNN[14] 算法对原始 R-CNN 进行了相当大的改进,但该模型仍然依赖于外部区域搜索算法。上述算法属于 two-stage 检测算法,这类算法需要先产生候选区域,再对 ROI(region of interest)做分类和位置预测。

由于校园接送系统对人脸检测的检测速度要求较高,本文采用的是基于 YOLOv3 的 one-stage 目标检测方法,用来检测视频帧中出现的人脸。

YOLOv3 只需要一个网络就可同时产生 ROI 并预测出物体的类别和位置坐标,是目前比较流行的端到端的目标检测算法。其基本思想是:首先对输入的图像进行卷积和一系列残差操作提取特征网络,产生 3 个尺寸像素为 $shape * shape$ 的特征层的输出,将其放入 YOLOv3 进行解码,接着对提取的 3 个特征层分别进行处理,输入图像分成 $shape * shape$ 个网格单元,每个网格单元负责预测其右下角区域的物体,若物体的中心点落在这个区域,这个物体的位置就由该网格点来确定,同时每个网格单元预测生成 3 个先验框,通过非极大值抑制算法排除冗余的候选框。由于 YOLOv3 中的 K-Means 聚类对初始聚类点的选择存在缺陷,通过随机选择 K 个点作为聚类中心,就导致假如初始点的位置选择不当,则最终的聚类结果会很糟糕或者需要进行多次随机初始化聚类中心才能得到良好的聚类结果。而 K-Means++算法对聚类中心的选择遵循距离聚类中心越远的点有更高的概率被选为下一个聚类中心的原则,因此 K-Means++算法能显著地减小分类结果的误差,对聚类中心的选取更加有效。

本文使用 K-Means++算法替换原始的 K-Means,对先验框聚类中心的位置选取进行改进,逐个中心点输入进行计算,大大降低了初始聚类点选取的影响。YOLOv3 网络架构如图 2 所示。图 2 中,YOLOv3 主干网络采用了 ResNet 思想的 Darknet-53 网络结构,包含了 52 个卷积层和一个全连接层,借鉴了 FPN[15] 架构,采用多个不同尺度的特征图来进行对象检测。从主干网络获得相对输入图像

的 8 倍、16 倍以及 32 倍下采样的特征图，进行 3 个尺度的预测。在第 79 层特征图做上采样与第 61 层特征图融合，经过几个卷积层后得到 16 倍下采样特征图，适合检测中等尺度的对象。第 91 层特征图再次上采样与第 36 层特征图融合，得到 8 倍下采样的特征图，适合检测小尺寸的对象，同时改用 sigmoid 和交叉熵函数对 softmax 损失函数进行改进，能够支持多标签对象的预测。

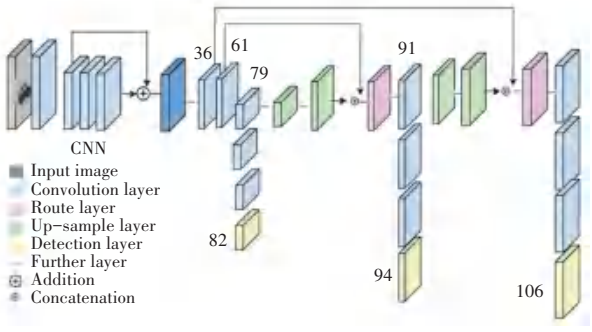


图 2　YOLOv3 网络架构图

Fig. 2　YOLOv3 network architecture diagram

由于本文的检测目标是人脸，因此选用香港中文大学发起的 WIDER FACE 人脸检测基准数据集训练人脸检测器，其图像是从可公开获得的 WIDER (Web Image Dataset for Event Recognition) 数据集中选择。本文选取了 32 203 张图像，并且还标记了 393 703张人脸，从中随机选择 70%、10%和 20%分别作为训练集、验证集和测试集。本文将 WIDER FACE 数据集标注好的人脸边框转成 VOC 格式，再进一步转换为 YOLO 标签，将检测类别修改为 face，同时修改配置文件使其使用于小目标检测，因为人脸基本都比较小。为了适应多尺度的目标检测，本文在训练开始前，首先进行实时数据增强的随机预处理：对训练数据进行数据增强，其本质是对图片进行平移缩放以及色域变换，使原始图片更加丰富多彩，训练的模型更具有鲁棒性。训练过程 *batch* 设置为 32，*subdivisions* 设置为 8，最大迭代次数约为 11 592 次，共 50 个 *Epoch*。在精确度相当的情况下，YOLOv3 算法检测速度更快，并在 WIDER FACE 数据集上评估本文模型的检测性能，对比结果见表 1。由表 1 分析可知，本文 YOLOv3-416 模型的 *mAP* 比原论文稍高。

本文利用训练好的 YOLOv3 人脸检测模型 yolov3_16000.weights，对多目标跟踪 MOT16 基准数据集[16]进行测试。MOT16 是 2016 年提出的多目标跟踪 MOT Challenge 系列的一个衡量多目标检测跟踪方法标准的数据集，包含训练集和测试集，本文选取了测试集中的 MOT16-11-raw.webm 和 MOT16-12-raw.webm 进行测试，显示每一帧检测的人脸数量，实验结果如图 3 所示。

表 1　本文与原论文 YOLOv3 模型的 *mAP* 对比

Tab. 1　Comparison of *mAP* between this paper and the original paper YOLOv3 model

Model	*mAP* - 50
YOLOv3-416(原论文)	55.3
YOLOv3-416(本文)	55.5

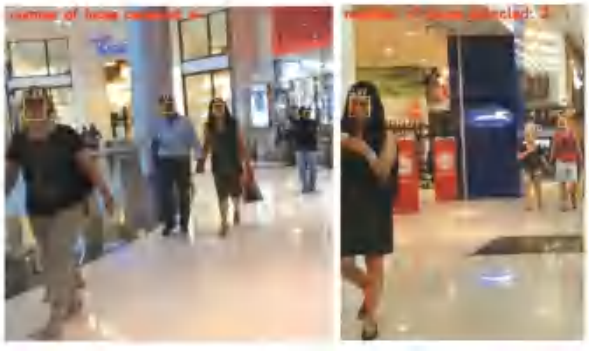


图 3　MOT16 测试结果

Fig. 3　MOT16 test results

同时，本文使用本地摄像头进行了测试，均能准确检测到人脸，测试结果如图 4 所示。

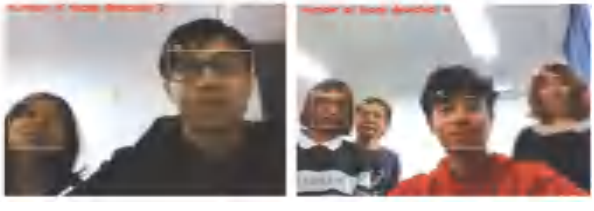


图 4　本地摄像头人脸检测测试结果

Fig. 4　Local camera face detection test results

2.2　人脸图像质量评价算法 FIQUE

基于视频的人脸识别，同一个人脸很可能会多次出现在视频中，而在重复出现的过程会呈现不同的人脸图像质量，那么选取合适的人脸进行识别是非常关键的。因此，本文使用评价算法 FIQUE 评估从多个人脸中筛选质量更好的人脸，从而避免了每拍摄到一个人脸就进行识别，能够提高系统的识别效率。

由于普通摄像头拍摄的视频帧往往面临一些复杂的现实环境，例如：人脸遮挡、光线不充足、人脸多、人脸姿态不正以及人脸尺寸过小等，这都会降低人脸图像的质量，进而影响到人脸识别算法的精度。因此，在人脸检测完成之后，本文提出一种在无参考图像的情况下，来预测图像质量分数的算法，将设定的阈值分数之外的人脸图像进行过滤，筛选掉不合

格的人脸,从而确保了输入到人脸识别环节中的人脸图像质量不会存在问题。

图像质量好坏是一个主观问题,针对如何为人脸图像给定质量得分,本文基于图像评估算法[17]的思想提出了一种人脸图像质量评价算法(Face Image Quality Evaluator, FIQUE),该算法的原理是对图像进行预处理后,从图像中提取均值减去对比度归一化系数,将该系数拟合成非对称广义高斯分布 AGGD[18]的结构,把提取到的拟合的高斯分布特征输入到支持向量机 SVM[19]中进行回归,进而得到图像质量的评估结果。下面给出人脸图像评价算法相关的系数和计算公式。

归一化系数 $I(x,y)$ 的定义如下:

$$\hat{I}(i,j)=\frac{I(i,j)-\mu(i,j)}{\sigma(i,j)+C} \tag{1}$$

$$\mu(i,j)=\sum_{k=-K}^{K}\sum_{l=-L}^{L}w_{k,l}I_{k,l}(i,j) \tag{2}$$

$$\sigma(i,j)=\sqrt{\sum_{k=-K}^{K}\sum_{l=-L}^{L}w_{k,l}\left(I_{k,l}(i,j)-\mu(i,j)\right)^2} \tag{3}$$

上述公式中,$\mu(i,j)$ 表示高斯滤波得到的结果,$\sigma(i,j)$ 表示标准差,该系数的优点在于对图像纹理等特征的依赖性较弱,这样提取出来的特征更具有普适性。广义高斯分布 GGD[20]的定义见式(4):

$$f(x;\alpha,\sigma^2)=\frac{\alpha}{2\beta\Gamma(1/\alpha)}\exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right) \tag{4}$$

广义高斯分布共有 2 个参数。其中,参数 α 代表分布的“形状”,也就是衰减的速率;参数 σ 代表方差。利用公式(4)可以方便地拟合归一化后人脸图像的分布,在广义高斯分布的基础上,进一步利用了非对称广义高斯分布的参数估计拟合人脸图像四个参数方向的归一化内积[18],其定义见式(5):

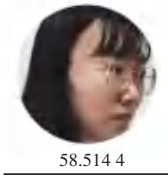
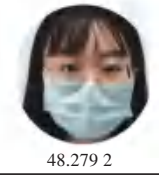
$$f(x;\alpha,\beta_l,\beta_r)=\begin{cases}\dfrac{\alpha}{(\beta_l+\beta_r)\Gamma(1/\alpha)}\exp\left(-\left(\dfrac{-x}{\beta_l}\right)^{\alpha}\right), & x<0\\ \dfrac{\alpha}{(\beta_l+\beta_r)\Gamma(1/\alpha)}\exp\left(-\left(\dfrac{x}{\beta_l}\right)^{\alpha}\right), & x\geq 0\end{cases} \tag{5}$$

非对称广义高斯分布一共有 3 个参数。其中,α 参数的含义同上,σ_l、σ_r 分别代表两侧的扩展速度,利用查找和匹配的方式,本文选择距离最小的值即为所求参数 α 的值。相比广义高斯分布,非对称广义高斯分布可以更好地拟合低质量图像产生的左右非对称现象。另外,本文将上述产生的这些特征向量输入到 SVM 算法中进行回归即可得到图像的质量得分。

本文选用 TID2008 图像质量评价数据集,其规定图像质量得分范围为 0~100,得分越小,主观图像质量越好。为了区分失真图像与自然图像,本文利用归一化系数对图像像素进行归一化,进而达到对人脸图像规范化的目的,使其特征向量被缩放至-1~1 之间,接着将特征向量输入到 SVM 和 LIBSVM 预测最终的质量分数。本文分别从表情、噪声、角度、模糊、遮挡维度,对人脸图像进行评估实验,各个维度对应的质量分数见表 2,根据实际情况设置图像质量得分阈值。

表 2 各个人脸维度以及对应的质量分数

Tab. 2 Each face dimension and the corresponding quality score

正向维度	表情维度	模糊维度
		
30.008 4	42.038 9	37.000 2
角度维度	噪声维度	遮挡维度
		
58.514 4	52.905 1	48.279 2

2.3 人脸识别算法

基于手工特征和传统机器学习技术的人脸识别方法存在部分人脸无法识别以及识别侧脸时精确度较低等问题,被深度神经网络技术所取代[21]。本文将 2.1 节中 YOLOv3 检测到的人脸图像送入 Inception-ResNet-v1[22]模型,该模型会对输入的人脸图像进行特征提取,最终通过比较特征向量进行人脸识别。

Inception-ResNet-v1 网络是在 Inception 模块中引入 ResNet 的残差结构,如图 5 所示,两者组合成为一个更优的网络,与原始 Inception 模块对比,增加了 shortcut 结构,并且在 add 操作前使用了线性的 1 * 1 卷积操作。ResNet 残差网络的模型能够通过增加相当的网络深度来提高训练准确率,其内部的残差块使用了捷径(shortcut connection),由于网络层数较深,因此使用跨越 3 个卷积层网络的残差块[8],同时采用了恒等映射(identity mapping),保证了反向传播更新参数时不会导致梯度消失,很大程

度上解决了当今深度神经网络的网络退化问题，使得越深的网络对于抽象特征的提取和网络性能更优。

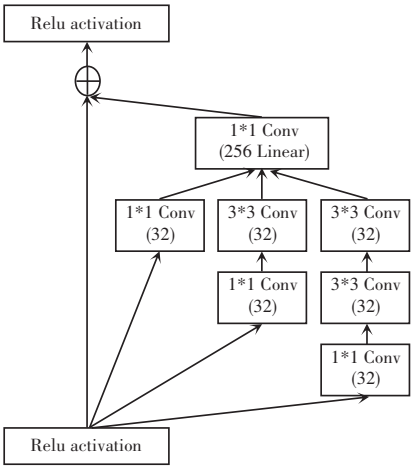


图 5　Inception-Resnet-A 模块

Fig. 5　Inception-ResNet-A module

为了提升人脸识别的速度，本文使用的 Inception 网络进行了卷积分解，将 5 * 5 卷积分解为 2 个 3 * 3 的卷积网络[8]，使得网络参数减少进而提升网络特征提取的能力，并且使用 ReLU 激活函数，使得人脸识别准确率更高。

3　实验结果与分析

3.1　实验数据与环境

本文的实验平台采用 Linux 操作系统 Ubuntu16.04，GTX1060Ti 显卡，3.2 GHz，16 GB 内存，256G 固态硬盘，CUDA 版本为 10.2，集成开发环境 IDE 使用 PyCharm，集成包使用的是 Anaconda3，Python 版本为 3.7.0，Tensorflow 版本为 1.14.0。为了测试本文的人脸识别模型的精确度，选用了下列人脸数据集进行测试。对此拟展开分述如下。

(1) LFW(Labled Faces in the Wild) 人脸数据库进行测试，主要用来研究非受限情况下的人脸识别问题，是测评人脸识别算法性能的重要指标。LFW 数据集提供的人脸图片均来源于生活中的自然场景，尤其由于对姿态、光照、表情等因素影响导致即使同一人的照片差异也很大，其共有13 233张人脸图像，每张图像均给出对应的人名，共有5 749个人。

本文从中随机选择了6 000对人脸组成了人脸验证 lfw_pairs.txt 文件，其中3 000对是属于同一标签的 2 张人脸图像，另外3 000对属于不同标签的人脸图像。本文使用了13 175张人脸图像，每次测试选用正负样本各 300 对，重复 10 组测试，即通过记录6 000次人脸测试结果的系统答案与真实答案的比值计算人脸识别的准确率。

(2) CFP(Celebrities in Frontal-Profile)[23] 数据集，主要是用来检测野外环境下的人脸识别。该数据集包括了 500 个类别 ID 的人脸图像，每个 ID 均有 10 张正面人脸照片和 4 张侧面人脸照片。CFP 数据集分为 2 个部分，分别是 frontal-frontal 和 frontal-profile。其中，前者是从正脸图像中选取的，后者的验证对象则是正面、侧面人脸各一张组成。本文测试集包含来自同一类别和不同类别的各 350 张人脸图像。

3.2　实验结果分析

为评估本文算法模型的精度，与经典的 Dlib 人脸识别库进行比较，两者在 LFW 数据集的识别准确率对比见表 3。由表 3 分析可知，本文算法的识别率稍好于 Dlib。

表 3　本文算法与 Dlib 的识别准确率

Tab. 3　The recognition accuracy of the algorithm in this paper and Dlib

算法	骨干网络	识别率/%
Dlib	ResNet34	98.40[24]
Ours	Inception-ResNet-v1	98.57

实验还利用本文算法在 AgeDB30 和 CFP 数据集进行测试，最终识别结果见表 4。

表 4　本文算法在不同数据集识别准确率

Tab. 4　Recognition accuracy of the proposed algorithm in different datasets

数据集	骨干网络	识别率/%
AgeDB-30	Inception-ResNet-v1	97.18
CFP-FP	Inception-ResNet-v1	96.01
CFP-FF	Inception-ResNet-v1	98.52

4　结束语

本文研究的是基于视频的人脸识别的应用，在不需要增加额外硬件设备的前提下，利用校园普遍具备的摄像头，很好地解决了现阶段低年级学生上下学接送环节的安全隐患。

针对接送过程中人流量大、人脸重复出现以及运动中人脸姿态变化等问题，为了对视频帧中出现的所有正向完整人脸进行有效的检测，本文选用基于 YOLOv3 的人脸检测算法，实验结果也表明了本文的人脸检测模型具有可靠效果。接着对检测到的人脸使用 FIQUE 算法进行人脸图像质量评价，筛选质量分数合格的人脸图像，保证人脸识别环节的人

(下转第 12 页)