DOI:10.19789/j.1004-9398.2018.05.006

基于深度学习的视频会议画面人脸光照对称性评价方法

王 珅[1] 张杰良[2] 张洪英[3]

(1. 清华大学电子工程系,北京 100084; 2. 中国人民解放军 61416 部队,北京 100036;
3. 中国人民解放军 31006 部队,北京 100036)

摘 要

人脸光照对称性是影响视频会议画面人脸质量的一个关键因素,本文通过分析人脸光照特性,引入卷积神经网络,提出一种基于深度学习的视频会议画面人脸光照对称性评价方法,在实验的基础上验证了与主观标记的一致性.

关键词: 视频会议,人脸光照对称性,卷积神经网络,深度学习.
中图分类号: TN911.73

0 前 言

视频会议画面的特点是侧重于人物的特写,即主要拍摄人物的正面画面,针对这一应用特点,视频会议画面质量的评价主要侧重于人物脸部画面质量的评价,主要包括脸部画面是否居中对称、光照是否对称均匀、对比度和清晰度是否合理. 在视频会议保障过程中,为了使脸部画面更为清晰可见,视频会议室的光照通常采用人工光源进行补光,在补光过程中需要保证光照既不能过量,也不能过低,而且光照强度适中,同时还要均匀,以免影响画面质量. 本文针对视频会议参与人员的近景特写图像,提出并实现了一种基于深度学习的人脸光照对称性评价方法. 该方法基本思想在于采用目前流行的深度学习模型: 首先大量收集会议室相关图片及视频样本,使数据有足够的代表性,能涵盖在视频会议画面中人脸光照对称性出现的各种情况; 其次使用有经验的人对这些样本的质量进行主观评价,使模型可以反映人的主观评价; 最后建立模仿人类的神经网络的深度学习模型,使得最终的模型与人的评价大致符合,其基本框架如图 1 所示.

1 深度学习 CNN 模型简介

深度学习(deep learning)用来解决抽象认知的难题,将人工智能这一领域带上了一个新的台阶. 区别于传统的浅层学习,深度学习的不同在于:

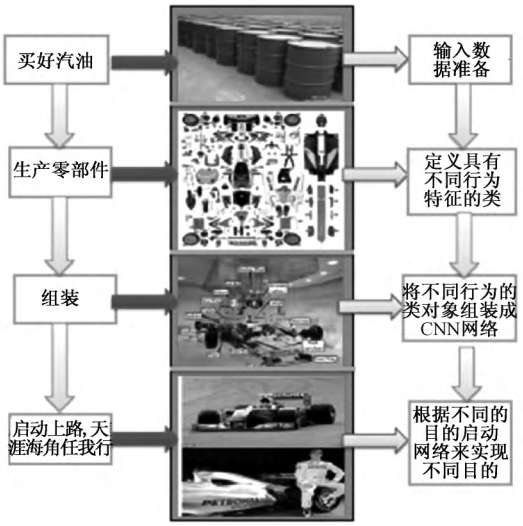


图 1 深度学习基本框架

1) 强调了模型结构的深度,通常有 5 层、6 层,甚至 10 多层的隐层节点;

2) 明确突出了特征学习的重要性,也就是说,通过逐层特征变换,将样本在原空间的特征表示变换到一个新特征空间,从而使分类或预测更加容易.

卷积神经网络(convolutional neural networks,简称 CNN)是一种基于多层神经网络的深度学习模型,采用反向传播算法 BP(back propagation)设计并进行训练. 相对于其他深度学习模型,在处理多维数据时 CNN 效果更加显著,识别错误率低,可应用于图像、视频的检测、识别和评价领域[1-3].

收稿日期: 2017-10-16

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

卷积神经网络通常至少有 2 个非线性可训练的卷积层，2 个非线性的固定卷积层(又叫 pooling layer) 和 1 个全连接层，一共至少 5 个隐含层. CNN 的结构受到著名的 Hubel-Wiesel 生物视觉模型的启发，尤其是模拟视觉皮层 V1 和 V2 层中 simple cell 和 complex cell 的行为. Hinton 和他的学生在著名的 ImageNet 问题[4] 上通过采用更深的 CNN，使得图像识别大踏步前进. 在 Hinton 的模型里，输入图像的像素，没有用到任何的人工特征，特征都是学习出来的. 与人工规则构造特征的方法相比，利用大数据来学习特征，更能够刻画数据的丰富内在信息. Hinton 深度学习 CNN 模型如图 2 所示.

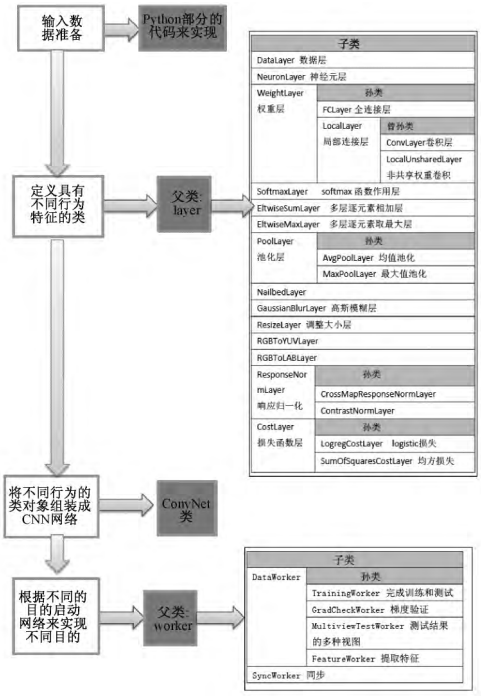


图 2　Hinton 深度学习 CNN 模型

2　人脸光照对称性检测算法

2.1　人脸光照对称性评价基本方法

一般来说，在理想状况下，人脸基本是左右对称的，但由于某些外界因素的出现，如人脸区域出现非均匀光照以及非正面的人脸姿态，往往会影响人脸的左右对称性. 因此，人脸的左右对称性可以用来当作评估人脸图像失真中的非均匀光照失真和非正面姿态失真的评价指标. 人脸图像左右部分的差异越小，则人脸越对称，失真程度越小; 反之，左右部分的差异越大，则左右半脸越不对称，失真程度就越大.

由于人脸的五官与肤色有明显的差异，为了避免五官对亮度分布的影响，需要提取脸部、眼睛、鼻孔和嘴等. 从算法角度出发，首先检测到画面中的人脸，之后提取人脸中的关键点，如五官的位置，人脸的轮廓等，最后在剔除五官之后，评价剩余人脸图像的整体对称性.

本文主要展开讨论人脸检测算法和眼睛检测算法，鼻孔、嘴巴等其他五官的检测算法类似于眼睛检测算法.

2.2　人脸检测算法

为了实现对会场人脸图像的质量评价，首先我们必须检测定位到人脸的具体位置，因此，首先我们需要实现人脸检测功能.

我们采用了基于多视角多通道 boosting 的人脸检测方案. 在人脸检测模块中，首先将一幅输入图像分解成许多不同位置、不同尺度的子窗口图像; 然后由分类器判断每一个子窗口图片是不是人脸; 最后，合并所有被判定为人脸的窗口和其附近的窗口以形成最终检测结果，如图 3 所示.

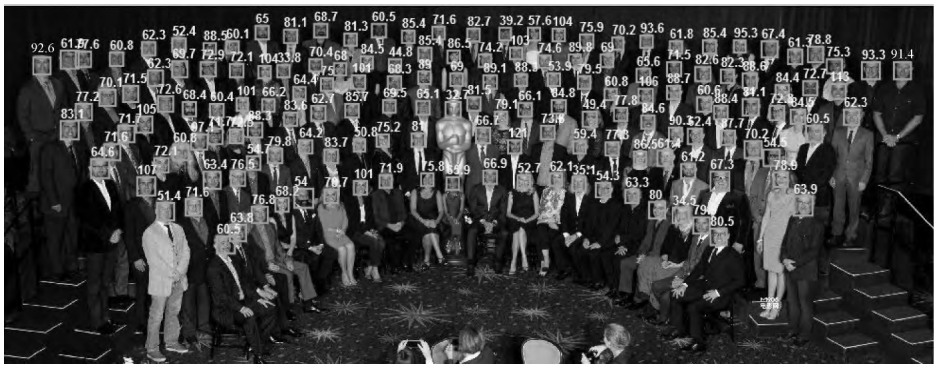


图 3　多视角多通道 boosting 人脸检测方案

人脸级联分类器的正样本训练集是 178 000 张尺寸为 20 × 20 像素的各种类型人脸图像. 候选的弱分类器通过在多尺度 Haar 特征上的阀值调整来构造. 每次迭代过程中，AdaBoost 学习算法首先计算出

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

每一个候选特征的最优阈值,这个阈值使特征所代表的弱分类器在训练集上加权分类错误率最小,然后在这些最优加权分类错误率中选择最小的那个弱分类器作为最优分类器. 当训练好一个强分类器后,后续级联分类器中的强分类器使用重采样的负样本,它们来自于已经训练好的级联强分类器.

2.3 眼睛检测算法

我们采用排序学习的方法实现眼睛定位和特征点对齐. 眼睛定位也可以称作眼睛检测,实现方法类似于人脸检测,并使用基于扩展 Haar 特征的级联 AdaBoost 分类器进行学习,但是被检测的区域位于检测到的人脸上半部分,并需要输出眼睛位置.

不戴眼镜的眼睛检测是一个相对容易的工作. 而在人脸图像中,在正面可见光情况下,戴眼镜的眼睛检测更加困难. 这是因为眼镜对光线的镜面反射引起眼睛区域的变化比可见光图像更加剧烈. 这导致直接使用 Haar 特征加 AdaBoost 的级联解决方案不能满足眼睛检测的精度和速度要求.

因此,我们使用从粗到精、分而治之的方法克服这个问题,人脸和眼睛检测的设计结构如图 4 所示.

该结构包括一系列的递增使用特殊眼睛数据训练的复杂眼睛检测器. 当检测到人脸图像后,通过一个粗糙眼睛检测器来分别检测左、右眼睛,这个检测器使用所有左眼或右眼的样本来训练. 这样可以检测到几乎所有的眼睛子窗口并且拒绝掉 95% 的非眼睛子窗口. 接下来,一个精细的不带眼睛的检测器将会检查每一个粗糙的眼睛子窗口,并检验它是否属于不戴眼镜的模式. 假如检测成功,将会返回一个眼睛区域; 如果失败,所有子窗口将会通过一个精细的戴眼镜的检测器,并检验它是否是一个戴眼镜的模式. 通过合并这些检测结果将会得到最终决策. 这种设计结构使快速、准确的戴眼镜(不戴眼镜)人脸和眼睛检测变为可能,如图 5 所示.

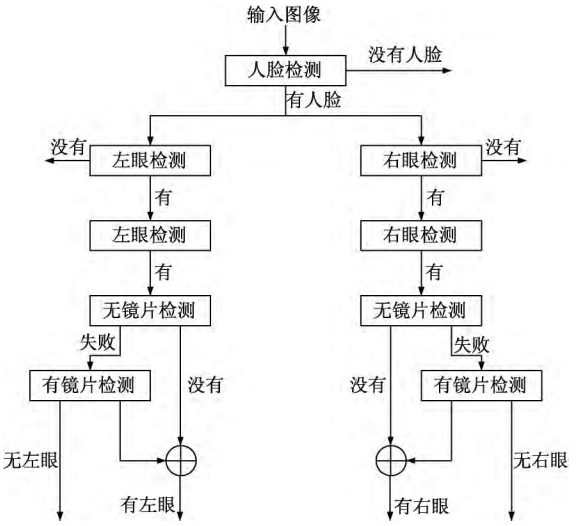


图 4 人脸检测流程



图 5 眼睛检测

这种设计主要考虑以下几个方面: 通过对戴眼镜近红外图像特征的观察,我们可以看到(1)戴眼镜和不戴眼镜的图像显示出巨大差异; (2)因为镜面反射影响更趋近于人脸的内侧部分而不是眼睛的中心,所以戴眼镜的两只眼睛图像不是左右对称的. 混合使用戴眼镜、不戴眼镜和左、右眼图像将会增加样本的变化程度. 然而,由于不戴眼镜的检测器已经十分准确、高效,并且我们已经大体知道左、右眼睛的先验位置,所以我们手动把眼睛图像分为三个子集: 不戴眼镜的眼睛图像、戴眼镜的左眼图像和戴眼镜的右眼图像,并且为每一类训练一个有效的眼睛检测器. 因为每一个子集样本的变化相应减少,显著提高了最终检测器的性能.

3 评价方法流程

基于深度学习的视频会议画面人脸光照对称性评价方法的流程如图 6 所示,包括以下具体步骤:

1) 输入一张图片,检测到图像中的人脸位置;

2) 对检测到的人脸进行关键点定位、姿态校正和光照预处理,得到归一化后人脸图像;

3) 根据人脸关键点,切取不同尺度,包含不同区域的人脸图像块,目前主要采取眼睛下方区域,

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

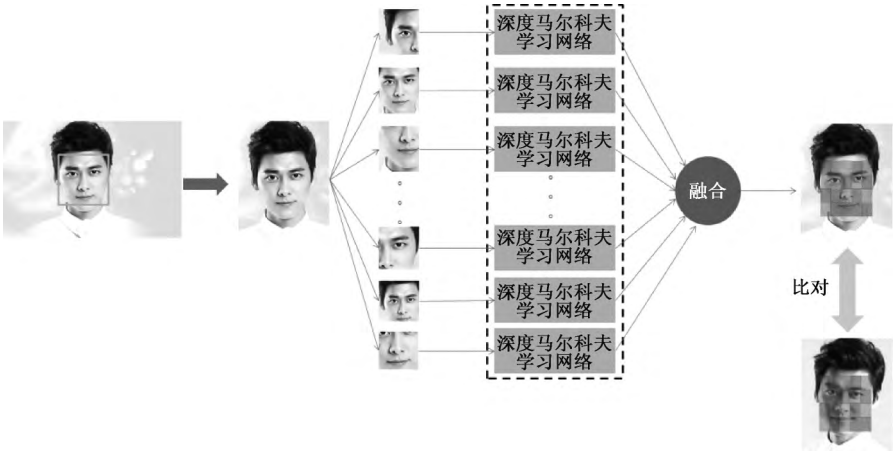


图 6　评价方法流程

额头区域,因为这些区域相对平坦,而且不受毛发的影响,整体亮度平均;

4) 采用马尔科夫随机场方法计算不同人脸图像特征判别向量间的一致性,得到对称性分数.

4　实验与结果分析

4.1　数据收集

参考国际上通用的研究人脸识别的数据库 YaleBExt [5],按照其建设标准,我们考虑了人脸的 5 种常见姿势(包括仰视、俯视、平视、左视、右视),组织 20 个人分别按照戴眼镜和不戴眼镜,在 10 种光照模式下共收集了 2 000 幅人脸画面图像.

在人脸画面图像主观评价过程中,采取相关专业人员按照 5 分制标准进行评价. 人工评价员根据参考图像,结合自己的主观判断,针对收集到的人脸画面图像,给出每个人针对画面的评价结果. 为了避免单个人的主观错误对整体识别效果的影响,采用多人同时评价同一画面. 综合同一画面不同人的评价得分,得出该画面的最终得分.

4.2　实验结果

视频质量专家组(video quality expert group,简称 VQEG) [6] 定义了四个指标,即 SROCC、KROCC、PLCC 和 RMSE,可用来检验客观结果和主观结果之间的一致性. 其中,SROCC 指标为斯皮尔曼等级次序相关系数,KROCC 指标为肯德尔等级次序相关系数,这两个指标代表评价方法的单调性,单调性强表明评价方法性能好,因此这两个指标的值需要接近于 1; PLCC 指标为皮尔逊线性相关系数,该指标代表评价方法的准确性,准确性高表明评价方法优,因此该指标的值同样需要接近于 1; RMSE 指标为均方根误差,该指标代表评价方法的误差,误差小表明评价方法性能准,因此该指标的值相对于客观评价值要尽量小.

实验过程中需要首先训练 CNN 深度学习模型,从 2 000 幅人脸画面图像中取出 1 800 幅图像对模型进行训练,然后对剩余的 200 幅图像进行评价,将获取的客观评价结果与主观标记分数进行数学分析对比,验证本文所实现方法的准确性、一致性. 一致性指标测试结果如表 1 所示.

表 1　一致性指标测试结果

指标	SROCC	KROCC	PLCC	RMSE
一致性	0.981	0.963	0.943	0.175

从表 1 中测试结果可以看出,指标 SROCC、KROCC 和 PLCC 都非常接近于 1,表明本文基于深度学习的评价方法具有相当高的准确性和单调性,同时 RMSE 值也比较小,说明该评价方法的误差也相对较小,证明本文实现的基于深度学习的评价方法与人工主观标记结果具有较高的一致性.

客观评价结果与主观标记分数之间的散点关系

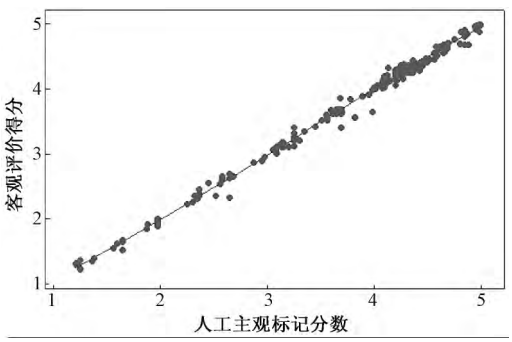


图 7　主客观评价测试结果散点图

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

图如图 7 所示,其中纵坐标表示人工主观标记的分数. 在散点图中,通过拟合曲线可以查看评价方法的一致性,如果拟合曲线平滑,而且各散点距离拟合曲线的距离紧凑,则表明评价方法的一致性高. 从图中可以看出,本文所实现的基于深度学习的评价方法的测试结果与人工主观标记分数具有较好的线性关系,一致性较高.

5 小 结

本文提出并实现了一种基于深度学习的视频会议画面人脸光照对称性评价方法,综合考虑了视频会议室人物特写画面时的各种因素,通过实验表明训练后的模型与主观评价方法一致,可用于自动高效地评价视频会议的人物图像质量.

参 考 文 献

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

[2] 乔林,张雄伟,史海宁,等. 深度学习应用中的常见模型 [J]. 军事通信技术,2016,37(1):98 – 104.

[3] 孙志军,薛磊,许阳明,等. 深度学习研究综述 [J]. 计算机应用研究,2012,29(8):2806 – 2810.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]. Lake Tahoe, Nevada, USA: Neural Information Processing Systems(NIPS), 2012: 1106 – 1114.

[5] Georghiades A S, Belhumeur P N, Kriegman D J. From few to many: illumination cone models for face recognition under variable lighting and pose [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(6): 643 – 660.

[6] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment [EB/OL]. (2005-01-10) [2017-06-15]. http://www.vqeg.org.

A Lighting Symmetry Assessment Approach of Facial Image for Video Conference Based on Deep Learning

Wang Shen[1] Zhang Jieliang[2] Zhang Hongying[3]

(1. Department of Electronic Engineering, Tsinghua University, Beijing 100084;
2. Unit 61416 of PLA, Beijing 100036; 3. Unit 31006 of PLA, Beijing 100036)

Abstract

Lighting symmetry of facial image was an important index of video conference image quality. This paper analyzed the lighting symmetry of human facial, introduced convolutional neural network, then proposed a lighting symmetry assessment approach of facial image for video conference based on deep learning. Experimental results showed that the approach was an effective approach for human facial image quality assessment, whose objective quality scores had better consistency with the subjective quality scores.

Key words: video conference, facial lighting symmetry, convolutional neural network, deep learning.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net