# 基于机器视觉的面部表情和属性识别研究

林中岚 秦运柏 刘俨 何润卓 李雨晴
（广西师范大学电子与信息工程学院 广西 桂林 541004）

**摘要：**人脸面部表情和属性识别是计算机视觉领域一个富有实用价值的研究课题。本研究聚焦于卷积神经网络的多任务学习，用于实现人脸识别及人脸属性（如年龄、性别、种族）分类。探讨了几种基于卷积神经网络的模型，包括 MobileNet、EfficientNet 和 RexNet。经过实验证明，这些模型在 UTKFace 数据集上实现了年龄、性别和种族的准确识别，同时在 AffectNet 数据集上进行情感分类方面也达到了先进水平。此外，在 AFEW 和 VGAF 数据集上，将经过充分训练的模型应用于视频帧中的面部区域特征提取，其准确率提高了 4.5%。

**关键词：**图像和视频处理；面部表情识别；年龄 / 性别 / 种族分类；多任务学习

## 0 引言

随着现代智能系统的不断发展，对图像和视频进行面部分析变得日益重要 [1]。这种分析涵盖了广泛的任务，包括对人脸年龄、性别、种族和情绪等特征的预测 [2,3]。近年来，基于深度卷积神经网络（CNN）的技术和模型取得了显著的进展，研究者们不断地提出各种创新性方法和模型，每年都涌现出新的模型和新的应用场景。这些研究不仅加深了对面部分析的理解，也在各个领域引发了广泛的兴趣。

然而，尽管这些复杂的 CNN 模型在提升性能方面表现出色，但它们的高度复杂性使得在资源受限的移动应用 [4] 或边缘设备上应用变得具有挑战性。因此，研究人员开始探索简化且易于实施的解决方案。这些解决方案中，多任务学习及按顺序训练不同问题的模型逐渐受到关注，因为它们不仅能够保持先进的性能水平，还能够避免繁琐的微调步骤，适用于各种任务和数据集 [3,5]。

在这个背景下，本文的主要创新点在于简化了训练流程，为各种面部分析任务提供了轻量级但高精度的 CNN 模型。通过在大规模的面部数据集上进行预训练 [6]，建立了一个通用的基础模型。与传统方法不同，建议利用人脸检测器输出的精确区域来对经过精心裁剪的人脸图像进行分类，从而避免了额外的边界信息。虽然在更小的面部区域上进行训练可能会稍微降低人脸识别的质量，但通过微调模型，能够实现更精确的人种分类和情感识别。此外，本研究还为基于视频的情感识别提供了新的可能性，通过提取网络特征，使单一模型在这一领域中取得了先进的结果。这些结果不仅对学术界有着重要的贡献，也为产业界和社会带来了潜在的影响。

## 1 研究背景与意义

在情感分类的静态图像任务中，AffectNet 数据集往往提供了最佳的结果 [3]。利用大型的 VGG-16 网络构建的超分辨率金字塔（PSR）[7]、深度注意力中心损失（DACL）[6] 等技术，都取得了极高的准确率。另一方面，在基于视频的情感识别中，通过 AFEW（静态面部表情）[3] 数据集进行评估。DenseNet 被用于从每帧图像中提取多个面部特征 [5]，同时结合 VGG13、VGG16 和 ResNet 的模型集合也取得了更好的结果 [8]。

在面部属性识别方面，如年龄、性别、种族等，通常使用在 IMDB-Wiki 或 UTKFace[2] 数据集上进行训练的 CNN 模型。目前，MobileNet v2（Agegendernet）、FaceNet、ResNet-50 及 Deep Expectation（DEX）的 VGG16 网络 [7] 等模型得到了广泛应用。虽然 DEX 的 VGG16 网络并未采用大规模人脸识别数据集进行预训练，然而多篇文献已经明确证明了这种预训练策略的优势 [6,8]。预训练不仅增强了面部处理任务之间的相似性，也提升了任务的整体性能。

在面部分析的多任务方法中，已经涌现了许多研究成果。例如，基于多任务学习的自我监督联合训

练在情感识别方面表现出色 [7]。PAENet[9] 是一个多任务网络，可以同时处理人脸识别、性别识别和面部表情理解，并且具有对新任务的快速适应能力。然而，值得注意的是，大多数效果最佳的 CNN 模型运行时间较长，难以适应很多实际应用的需求 [4]。因此，本文将重点放在轻量级的 CNN 架构上，如 EfficientNet 和 RexNet，以在性能和实际应用之间取得平衡。这种方案将有望在面部分析领域取得显著的进展。

## 2 基于机器视觉的面部表情和属性识别

### 2.1 多任务网络

本研究借助多任务神经网络 [8]，致力于解决多个人脸属性识别问题（图 1）。首先在超大型 VGGFace2 数据集 [10] 上对基础 CNN 进行了人脸识别预训练。尽管传统方法倾向于采用每张照片的中心裁剪方式来处理 224×224 区域（图 2a、图 2c），然而本文强调了采用多任务 CNN（MTCNN）进行人脸检测的优势（图 2b、图 2d），这样可以获得更高的质量，而且无需引入任何额外的边距。

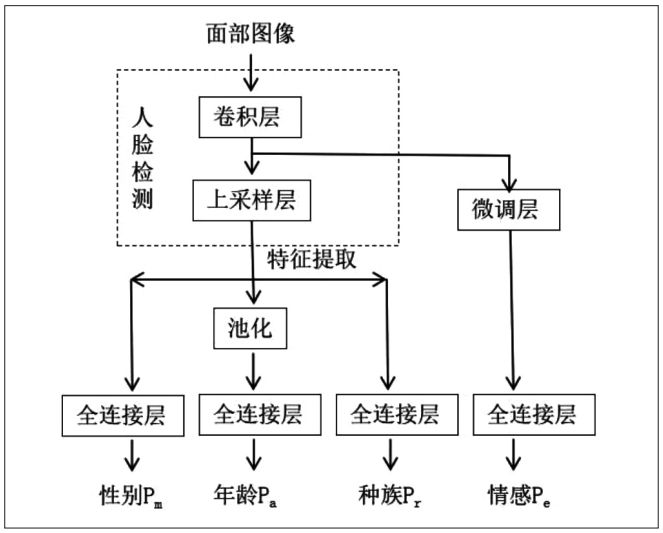
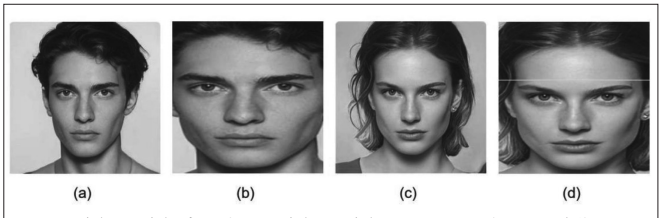



图 1 多任务面部表情和属性识别方法

鉴于本研究关注于轻量级 CNN，选择了 MobileNet、EfficientNet 和 RexNet 等架构，作为人脸识别网络的 Backbone。这些网络能够从中提取适于区分不同个体的面部特征 $\boldsymbol{x}$。这些特征可以用来预测稳定属性，如性别和种族，仅需使用简单的分类器，即全连接（FC）层。然而，与同一受试者的年龄通常变化缓慢不同，年龄是在不同人之间迅速变化的属性。因此，由图 1 可知，在特征向量 $\boldsymbol{x}$ 上增加了一个层，以便在最终的 FC 层之前预测年龄。尽管年龄预测是回归问题的一种特例，但在本文中，将其视为一种多类分类问题，需要预测被观察者的年龄是 1 岁、2 岁，直至 $C_a$ 岁 [8]。




（a）、（c）中心裁剪；（b）、（d）用 MTCNN 裁剪，无边缘

图 2 LFW 数据集中的图像样本

然而，许多其他面部属性会迅速变化，因此面部识别的特征需要与这些变化保持一致。以情感识别任务为例，即使面部表情相同，同一人在不同情感状态下的身份特征之间的类间距离应远小于不同个体之间的类内距离。因此，本研究认为，通过识别任务训练的 CNN 提取的面部特征不适用于直接用于情感识别。与在与人脸无关的数据集上（如 ImageNet）进行预训练的 CNN 相比，这种 CNN 更适合后者的任务，因为其低层特征更能够捕捉到边缘和角落等特征。因此，在情感数据集上对人脸识别 CNN 进行微调，以更好地利用有价值的面部特征信息，或用于预测与个体身份无关的面部属性。

### 2.2 训练流程

为了精简训练流程，采用逐步调优的策略，从人脸识别问题起步，逐渐拓展至不同的人脸属性识别任务 [11]。

首先，在 VGGFace2 数据集上对人脸识别 CNN 进行精心训练。该训练集涵盖了 300356 张照片，而测试集包含了额外的 223146 张图像。在预先在 ImageNet 上训练过的网络中引入了新的 Head，其中包含 8541 个输出，并融合了 softmax 激活函数。在一个 epoch 的训练周期内，将基础网络的权重固定，专注于头部层的学习。通过现代的锐度感知最小化（SAM）[6] 和学习率设置为 0.001 的 Adam 优化器，对分类交叉熵损失函数进行了精细优化。

其次，以相同的方法对整个 CNN 进行了 10 次训练，但此时将学习率调整为 0.0001。值得一提的是，本研究其余部分基于在验证集上表现出色的模型，包括 MobileNet-v1、EfficientNet-B0 和 RexNet-150，

它们的准确率分别达到了 90.3%、92% 和 93%。

再次，引入了专门用于预测年龄、性别和种族的独立头部，并对这些头部的权重进行了精心调整。训练数据集汇集了来自 IMDB-Wiki 数据集 [1] 的 30 万张正面裁剪面部图像，用于年龄和性别的预测 [8]。在优化传统的交叉熵损失函数时，将基础模型的权重冻结，仅关注新头部的权重更新。经过三轮的训练，基于 MobileNet 的模型在性别和年龄分类方面分别达到了 94% 和 11% 的验证准确率。

最后，为了实现最终的年龄预测，选择了 CNN 输出中后验概率最大的指数 $\{a_1,...,a_L\}$，其中 $L \in \{1,2,...,C_a\}$，并计算出它们的平均期望值 [8]：

$$\bar{a}=\frac{\Sigma_{l=1}^{L} a_l \cdot p_{al}}{\Sigma_{l=1}^{L} p_{al}} \quad (1)$$

在种族分类方面，专注于 UTKFace 数据集的一个子集，通过在训练过程中应用差异权重来提升不平衡类别的性能。该数据集的“对齐和裁剪人脸”部分涵盖了 22456 张常规图像，其中 20132 张用于训练，其余用于测试。在这个数据集中，涵盖了 $C_r$=5 个不同的种族类别。

进入最终阶段，对网络进行微调，以适应 AffectNet 数据集 [3] 中的情绪识别任务。该数据集的训练集由笔者精心准备，包含了 278521 张图像。其中，$C_e$=8 个情绪类别（中性、快乐、悲伤、惊讶、恐惧、愤怒、厌恶、轻蔑）汇聚成丰富多彩的情感画面。此外，该数据集涵盖了 7 种主要表情，唤起了情感世界的不同层面（排除了轻蔑这一情感）。正式的验证集则呈现了每个类别 500 张图像，分别对应 8 个和 7 个情感类别，总计 3700 张图像。在面对 7 种情感的分类问题上，探索了两种方法：首先，在经过削减的训练集上训练模型，集中精力探索情感世界的核心；其次，在包含 8 个情感类别的全体训练集上训练模型，虽然最终只使用了 7 个类别的分数（Softmax），但为了捕捉情感的微妙变化也同样不遗余力。无论是哪种情况，都运用了加权分类交叉熵（Cross Entropy Error）损失函数进行了精心的优化 [3]，以确保模型充分理解和表达情感世界的细腻之处。

在这个过程中，定义 $X$ 为训练图像，而 $y \in \{1,...,C_e\}$ 代表着其对应的情感类别标签。在此基础上，$N_y$ 则表达了第 $y$ 类训练实例的总数，而 $z_y$ 则是指示倒数第二层中的第 $y$ 个输出，在此处，运用了 *softmax* 函数进行软性的最大激活，如式（2）所示。

$$L(X,y)=-\log softmax(z_y)\cdot \max_{c\in\{1,...,C_e\}} N_c/N_y \quad (2)$$

与初始的人脸识别 CNN 预训练类似，借鉴了类似的训练方法。首要步骤是引入一个新颖的头部，带有 $C_e$ 个输出。在保持其余权重不变的同时，利用 SAM[4] 在 3 个 epoch 的时间内对这个新头部进行了针对性的训练。最终，在接下来的 10 个训练周期内对所有的权重进行了全面的调整。

### 2.3 基于视频的面部属性识别

基于视频的情感分类任务所涉及的数据集包含了用于训练分类器的精选视频片段。之前在 AffectNet 数据集上微调过的网络被运用于提取每一帧的特征。在每个帧中，选择了最大的面部区域，并将其输入到 CNN 中。倒数第二层输出的 $D$ 维向量被保存在帧描述符中，这一策略形成了视频描述符 [1]，其维度为 $4D$。对于 MobileNet 和 EfficientNet-B0 而言，分别是 4 × 1024=4096 和 4 × 1280=5120。这一计算方式通过将统计函数（平均值、最大值、最小值和标准偏差）[7,11] 应用于帧描述符来实现。接下来，借助 scikit-learn 库中的分类器，如 LinearSVC，或者进行了 1000 次的随机森林训练，通过 L2 规范损失函数对视频描述符进行分类处理。

面对多组视频情感识别任务，采用的解决方案与之类似，尽管每帧可能涵盖多个面部区域。因此，计算视频描述符的方式略有调整。首先，将单帧中所有面部情感特征的统计函数（平均值和标准偏差）融合在一起，构建了该帧的描述符。随后，使用相同的平均值和标准偏差函数对所有帧描述符进行聚合。这里并未采用最大和最小聚合函数，以保持最终描述符的维度。这一流程的巧妙设计使得能够更全面地把握视频中多组情感的变化和特性，从而更加准确地进行情感分类。

## 3 实验结果

### 3.1 人脸识别

在初次的实验中，选择了经典的人脸识别指标来评估 LFW（Labeled Faces in the Wild）数据集。具体而言，从 LFW 数据库中挑选了 $C$=565 个样本，这些样本在该数据库中至少拥有 2 张人脸照片。在

表 1 不同模型人脸识别的准确率 /%

| 卷积神经网络 | 中心裁剪 | WTCNN 裁剪，无边框 |
|---|---|---|
| SENet-50[10] | 97.1±2.9 | 96.6±2.2 |
| MobileNet | 90.8±3.6 | 92.2±3.5 |
| EfficientNet-B0 | 91.9±4.7 | 93.8±4.3 |
| RexNet-150 | 94.7±4.4 | 94.3±3.7 |

训练集中，只放入了 1 张人脸图像，其余的图片用于测试集。不同模型人脸识别的准确率见表 1，其清晰呈现了在 VGGFace2 数据集 [10,12] 上预先训练的多个 CNN 模型的人脸识别准确率及其标准偏差。这些准确率是通过进行 10 次随机的交叉验证估算获得的。

在人脸面部表情和状态识别任务中，基于传统 SENet 的人脸描述符 [10] 展现出更为精准的表现，特别是在每个样本只有一张训练图像的情况下，尤其是在对松散裁剪的人脸进行识别时。然而，对于经过无边界人脸检测器裁剪的人脸，其错误率增加了 0.5%。与此相反，通过对被检测到的人脸进行分类，而不是采用中心裁剪方法，所提出的描述符的准确率上升了 1% ~ 2%。因此，可以预见，针对不含背景的面部区域进行处理对于其他面部分析任务具有更为实际的价值。这个发现为未来的面部分析方法提供了有益的指导方向。

### 3.2 单张图像的面部表情识别

深入探讨了针对 AffectNet 数据集的情感分类模型。此外，还对第一训练阶段后的模型进行了分析，其中仅学习了分类头的权重，而其余部分保持与预训练的人脸识别 CNN 相同。这种策略允许基础网络提取适用于人脸识别的关键特征。同时，还对一些现有模型在 AffectNet 上进行了训练，并与本文提出的管道训练的模型进行了全面比较。这些模型采用了其他的预训练策略，包括在 ImageNet-1000 上预训练的 MobileNet、Inception 和 EfficientNet，以及在 VGGFace2 上预训练的 SENet[10]。将这些模型的性能与拟议管道训练的 CNN 进行了对比总结，这些数据都是在 AffectNet[3] 指定的训练集和验证集上进行的。

然后，使用在包含 8 个情感类别的完整 AffectNet 训练集上训练的模型来预测 7 个情绪类别的准确率稍低，但它的通用性更强，因为同一个模型可以同时应用于 8 种和 7 种情绪的预测。此外，实验支持了这样一种观点，即来自预训练 CNN 的身份特征并不适用于可靠的面部表情识别，尽管那些基于 MTCNN 裁剪的人脸进行训练的模型表现更为优越。值得强调的是，与 ImageNet 上预训练的 CNN 相比，所提出的方法进行训练的模型准确率有了显著的提升。即便是在 VGGFace2 数据集上预训练的 SENet 模型。在使用人脸检测程序时，选择预测的边界框而不添加任何边距是至关重要的。因此，基于 EfficientNet 的模型在 8 类和 7 类 AffectNet 情感分类任务中取得了显著的提高。

## 4 结语

本文引入了一种创新的训练流程，以在图像和视频的多个面部表情识别数据集中实现神经网络的卓越准确性。与现有模型相比，这种方法通过预先训练特征提取器，将庞大的 VGGFace2 数据集用于人脸识别，从而增强了人脸提取和对齐的鲁棒性。此外，利用了人脸检测器返回的裁剪过的人脸区域，无需额外添加边缘。因此，不仅取得了卓越的准确率，还保持了优异的速度和模型大小。

尽管通过训练获得的模型在面部表情特征方面表现出色，但本文仅采用了传统的分类器（如支持向量机、随机森林等）。因此，无法始终达到最先进方法的性能水平。未来的研究需要考虑采用更为复杂的分类器，如图卷积网络、变换器，以及帧 / 通道级的注意力机制，以进一步提升年龄、性别和情感识别的整体质量。通过引入这些先进的分类器，有望推动面部分析领域的性能进一步提升，为实际应用带来更多的价值。这也将为面部分析技术的发展开辟新的可能性。

## 参考文献：

[1] 王曼曼 . 基于生成式对抗网络的人脸属性识别对抗攻击研究 [D]. 南京：东南大学 , 2021.

[2] 和瑞丽 . 基于视觉的婴儿人脸属性和行为识别方法研究 [D]. 哈尔滨：哈尔滨工业大学 , 2020.

[3] 韩子阳 . 多面部姿态下的人脸表情识别 [D]. 苏州：苏州大学 , 2020.

[4] 王天宇 . 基于 PYNQ 平台的多人脸属性识别研究与实现 [D]. 南京：南京邮电大学 , 2022.

[5] 谭彬 , 杜炳德 , 赵雅琪 . 基于 Inception-V3 网络的多任务人脸属性识别研究 [J]. 无线互联科技 , 2022, 19(22):