高技术通讯 2024 年 第34 卷 第5 期:486-495

doi:10.3772/j.issn.1002-0470.2024.05.005

基于改进 HopeNet 的头部姿态估计方法①

张立国② 胡 林③

(燕山大学测试计量技术与仪器重点实验室 秦皇岛 066004)

摘 要 针对基于无需先验知识的头部姿态估计算法在复杂背景图像和多尺度图像场景下精度较差的问题,提出了一种基于改进 HopeNet 的头部姿态估计方法。首先在主干网络结构上增加特征融合结构使得模型能够充分利用网络的深层特征信息与浅层特征信息,提升模型的特征解析力;然后在主干网络的残差结构中增加特征压缩激励模块,使得网络能够自适应学习不同特征层重要程度的权重信息,让模型更加关注目标信息。实验结果表明,相较于 HopeNet,本文方法在 AFLW2000 数据集上精度提升了 31.15%,平均误差降到 4.20°,同时在复杂背景图像场景下有较好的鲁棒性。

关键词 头部姿态估计; HopeNet; 特征融合; 特征压缩激励; 自适应学习

头部姿态估计是指运用仪器设备采集数据后通过智能算法估计人体头部在三维空间下的旋转角信息。通常头部姿态空间旋转角信息用欧拉角表示为偏航角(Yaw)、俯仰角(Pitch)和滚转角(Roll)。头部姿态估计广泛应用在辅助驾驶、虚拟现实、面部分析等领域中[1-3]。如今移动设备如手机、笔记本和智能相机等电子设备的广泛使用,大幅降低了图像数据获取的成本,同时促进了基于图像的头部姿态估计算法的广泛应用[4-5]。其中传统基于红绿蓝(red green blue,RGB)图像的头部姿态估计方法分为 2 种:基于外观的方法和基于模型的方法[6]。基于外观的方法通常是假定获取的头部图像和头部姿态间存在某种映射关系,运用统计或者相似度计算的方法来推断头部姿态[7-9]。该方法原理简单,在实际场景中局限较大,例如在不使用插值等优化方法的情况下无法估计连续姿态。基于模型的方法借助面部几何信息或面部关键点信息与标准头部模型匹配得到头部姿态。虽然该方法能够获取到连续的头部姿态估计值,但姿态估计准确度严重依赖面部信息的精度。

随着硬件性能的提升,深度学习技术在众多领域得到了广泛的应用。其中基于深度学习的头部姿态估计也取得了一系列的研究成果。Patacchiola 等人[10]将卷积神经网络方法应用于头部姿态估计,并采用自适应梯度下降方法优化模型,但是该方法网络结构简单,精度不高。Zhou 等人[11]设计了端到端的训练模型,在一定程度上提升了网络精度,并采用全景图像数据作为训练集,实现水平旋转角360°的头部姿态估计,但是该方法在复杂背景图像上精度较差,且实际应用中全景图像采集困难。Berral-Soler 等人[12]为了平衡模型推理速度和精度,重新设计轻量化模型,实现模型在推理中达到实时的效果,但是该方法以牺牲精度为代价减少模型计算量。Ruiz 等人[13]提出了无需先验信息的头部姿态估计,将分类方法和回归方法相结合解决了低质量图像精度差的问题,但是该方法在多尺度图像上精度较差。

为了解决上述问题,本文在文献[13]的启发下提出了基于改进 HopeNet 的头部姿态估计方法。采用的优化策略如下:(1)在主干网络中设计特征融合模块,将深层特征与浅层特征融合,使得网络能够

① 河北省中央引导地方专项(199477141G)资助项目。
② 男,1978 年生,博士,副教授;研究方向:机器视觉,故障诊断,虚拟现实;E-mail: zlgtime@163.com。
③ 通信作者,E-mail: 13212818772@163.com。
(收稿日期:2023-02-28)

充分利用不同深度的特征信息;(2)在 ResNet50 的残差结构中增加通道压缩激励模块(squeeze-and-excitation,SE),使得网络能够自适应学习不同特征层对应的权重来区分网络层的重要程度;(3)在数据增强上采用随机遮挡、随机灰度变换、随机亮度变换和随机对比度变换策略可以有效防止模型过拟合、提升数据多样性和提升实际场景的鲁棒性。

1 无需先验知识的头部姿态估计模型原理与改进

无需先验信息的头部姿态估计方法[13]将头部姿态估计问题视为分类与回归相结合的问题。模型在推理时输入头部图像提取特征,然后将得到的特征用全连接计算得到分类结果,最后将分类结果进行积分运算得到最终头部姿态角。其网络结构如图 1所示,主要由特征提取和头部姿态回归 2 个部分构成。其中特征提取网络采用的是ResNet50,有

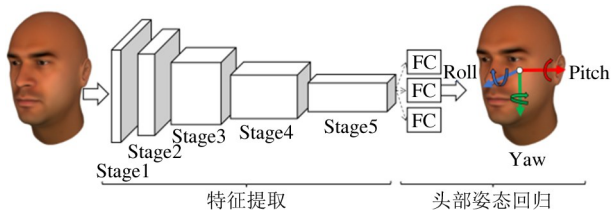


图 1 无需先验信息的头部姿态估计网络结构

49 个卷积层和 5 次下采样,去掉最后的池化层和全连接层。虽然主干网路采用残差网络结构解决了网络过深而导致梯度消失的问题,但是没有区分网络层的重要程度从而引入图像背景噪声信息。头部姿态回归部分仅仅利用特征网络深层网络特征信息,导致模型在多尺度图像上精度较差。针对上述问题本文在该方法基础上进行改进和优化。

本文提出的基于特征融合与压缩激励的头部姿态估计方法网络结构如图 2 所示,改进内容如下所述。

(1) 主干网络采用去掉最后池化层和全连接层的 ResNet50,然后添加特征融合结构(Fusion)。其计算过程可以描述为:先将 $S2$(Stage2)、$S3$(Stage3)和 $S4$(Stage4)的输出用不同卷积核大小的分组卷积运算得到 $S2^*$、$S3^*$ 和 $S4^*$,然后将 $S2^*$ 进行卷积下采样与 $S3^*$ 相加融合得到 $M2$,最后将 $M2$ 卷积下采样与 $S4^*$ 融合得到 $M3$。同理得到最终融合输出特征 *out*。通过特征融合计算实现深层特征与浅层特征的融合利用,提升模型在多尺度图像上的精度。

(2) 在主干网络 Stage1、Stage2 和 Stage3 阶段增加信道压缩激励模块,即在 ResNet50 网络的残差结构中增加通道压缩激励模块。该结构可以使网络学习特征层的重要程度增强从而强化关键层特征,弱化非关键层,让网络更加关注前景信息,有助于关

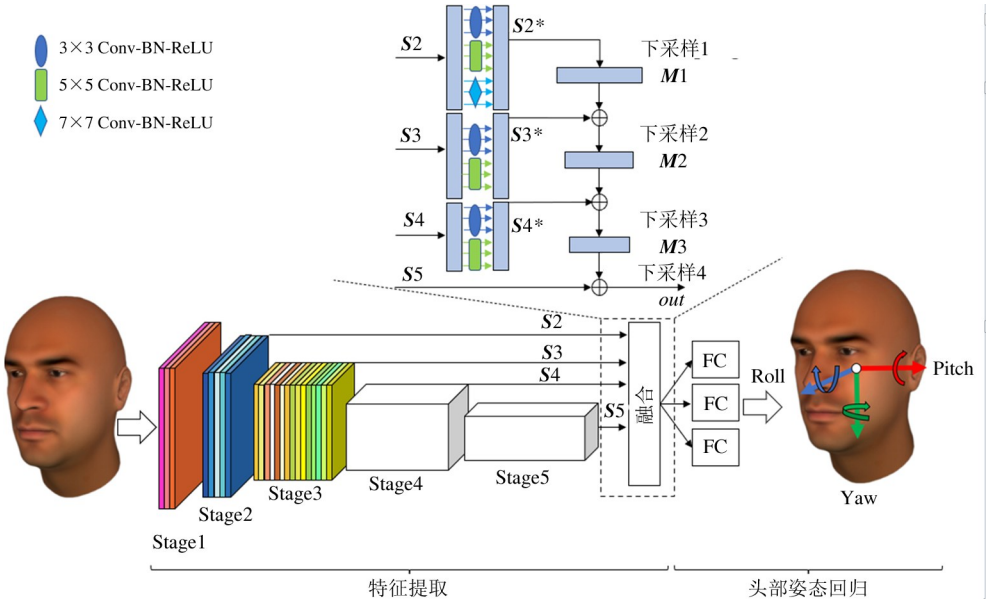


图 2 基于特征融合与压缩激励的头部姿态估计方法的网络结构

键信息向后传播。

2　通道压缩激励与特征融合

2.1　通道压缩激励

通道压缩激励通过构建特征图与特征通道的映射关系,实现自适应学习通道上的特征响应从而区分各个通道维度不同特征层的重要程度。普通残差网络结构和嵌入通道压缩激励的残差网络结构如图 3所示。其中虚线框所示为普通残差网络结构,计算时先将输入特征用卷积神经网络提取特征后与输入特征相加得到输出结果,计算过程可以表示为式(1)。

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} \oplus F_{\text{conv}}(\boldsymbol{X}) \tag{1}$$

式中, $\tilde{\boldsymbol{X}} \in R^{C\times H\times W}$ 表示残差网络输出结果; $\boldsymbol{X} \in R^{C\times H\times W}$ 表示网络输入;"⊕"表示两个特征图对应元素相加; F_{conv} 表示卷积计算。

如图 3 所示,在传统残差网络结构基础上增加改进的特征压缩激励结构。本文在传统压缩激励结构[14]基础上增加平均池化(average pool,Avgpool)和最大池化(max pool,Maxpool)2 个计算分支。在输出时,2 个分支特征进行融合得到输出结果。改进压缩激励模块的计算过程具体如下所述。首先,分别用 Avgpool 和 Maxpool 将输入特征 $\boldsymbol{S}_0$ 进行关键特征提取后得到2 个一维的特征 $\boldsymbol{S}_1$、$\boldsymbol{S}_2 \in R^{1\times 1\times C}$,该一维特征具有输入信息的全局特征。然后,将 2 个 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 分别用 2 个全连接和激活函数构成的多层感知机(multilayer perceptron,MLP)提取特征(r 为压缩率超参数,参数设置可参考文献[14]),得到 $\boldsymbol{S}_1^*$ 和 $\boldsymbol{S}_2^*$,该结构使得网络能够自适应地学习对应特征层的重要程度权重。最后,将提取到的权重特征信息 $\boldsymbol{S}_1^*$ 和 $\boldsymbol{S}_2^*$ 按元素相加融合后得到特征 $\boldsymbol{M}$,$\boldsymbol{M}$ 特征信息可用于调整和增强输入的特征图。

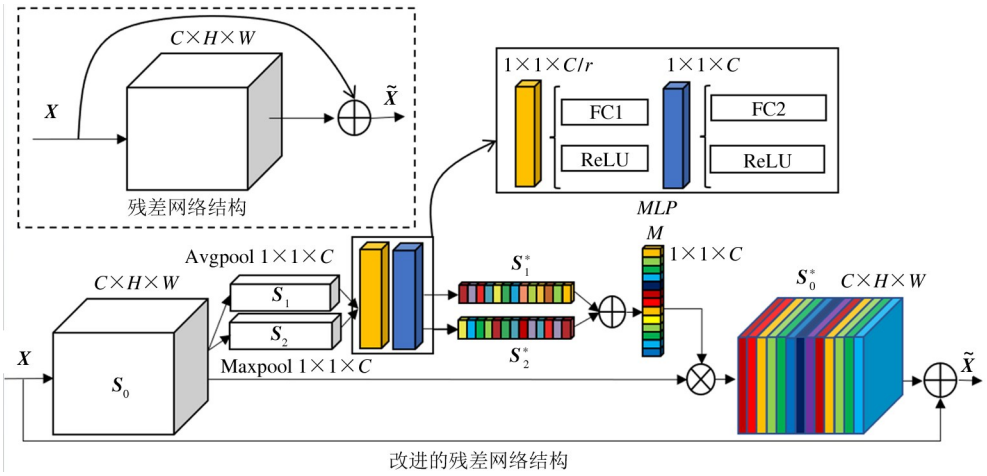


图 3　残差网络结构与增加改进信道压缩激励的残差网络结构

如图 3 所示,改进后的残差网络首先将输入数据 $\boldsymbol{X} \in R^{C\times H\times W}$ 采用卷积网络提取特征得到 $\boldsymbol{S}_0$;其次用特征压缩激励对 $\boldsymbol{S}_0$ 提取层方向上的全局特征信息 $\boldsymbol{M}$;然后用 $\boldsymbol{M}$ 元素分别与 $\boldsymbol{S}_0$ 在通道维度上相乘得到增强后的特征 $\boldsymbol{S}_0^*$。最后将输入特征 $\boldsymbol{X}$ 与 $\boldsymbol{S}_0^*$ 对应元素相加得到输出结果 $\tilde{\boldsymbol{X}}$。

残差网络计算过程和特征压缩激励计算过程分别如式(2)和式(3)所示。

$$\begin{cases} \tilde{\boldsymbol{X}} = (\boldsymbol{M} \otimes \boldsymbol{S}_0) \oplus \boldsymbol{X} \\ \boldsymbol{S}_0 = F_{\text{conv}}(\boldsymbol{X}) \end{cases} \tag{2}$$

$$\boldsymbol{M} = F_{\text{MLP}}\left[\frac{1}{H\times W}\sum_{i=0}^{H}\sum_{j=0}^{W} f_x(i,j)\right] \oplus F_{\text{MLP}}\left[\max_{i\in H,\, j\in W} f_x(i,j)\right] \tag{3}$$

式中,H、W 和 C 分别表示特征图的高、宽和通道数;"⊗"表示两个特征图对应元素相乘; $f_x(i,j)$ 表示输入特征图 $\boldsymbol{X}$ 在 x 通道 (i,j) 位置像素的索引;

F_{MLP} 表示对输入数采用多层感知机提取特征，F_{conv} 表示卷积网络提取特征。

2.2 信息融合与损失计算

特征融合模块与损失计算模块结构如图 4 所示。其中特征融合模块计算时先将输入特征采用分组卷积的方式实现不同卷积核大小提取特征信息，然后采用卷积网络下采样的方式将浅层特征图和深层特征信息进行融合输出，提升网络的特征解析力和不同尺度图像的估计精度。损失计算部分特点是同时使用分类损失和回归损失 2 个监督信号进行反向梯度下降优化，从而提升网络训练速度和精度。

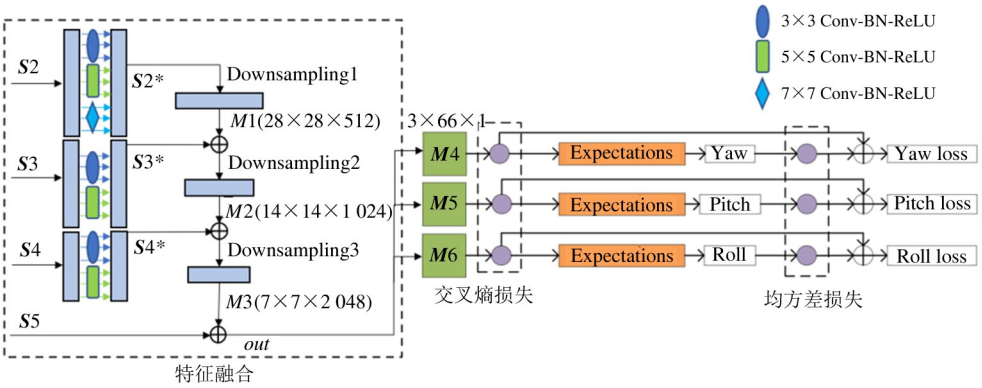


图 4 特征融合与损失计算结构

特征融合模块的具体结构如图 4 中虚线框部分所示。其计算过程是先将 ResNet50 提取的特征 $S2$、$S3$ 和 $S4$ 通过分组卷积计算得到 $S2^*$、$S3^*$ 和 $S4^*$。其中分组卷积计算时首先根据卷积核的个数在特征通道方向上均分特征。以 $S2$ 计算为例，一共有 3 种类型卷积核，首先将特征图在通道上平均分为 3 组特征 $\tilde{X}(h,w,c_1)$，$\tilde{X}(h,w,c_2)$ 和 $\tilde{X}(h,w,c_3)$；然后分别用卷积计算后的特征在通道上拼接得到 $S2^*$，卷积核大小分别为 3、5 和 7。特征边缘填充(Padding)大小计算方式为 K(kernel)/2 向下取整，在本实例中 Padding 分别为 1、2、3，步距大小为 1。分组卷积的计算过程如式(4)所示，分组卷积输出结果计算后如式(5)所示，式中 $\hat{K}^t$ 表示第 t 个卷积计算的卷积核大小，$\hat{Y}^t$ 表示第 t 组特征经过卷积计算后的结果。

$$\hat{Y}^t = \hat{X}(h,w,c_t) \cdot \hat{K}^t \tag{4}$$

$$S2^* = \text{Concat}(\hat{Y}^1, \hat{Y}^2, \hat{Y}^3) \tag{5}$$

同理计算得到 $S3^*$ 和 $S4^*$。然后将 $S2^*$ 用卷积核大小为 3、Padding 为 1、步距为 2 的卷积计算用于 2 倍下采样得到 $M1$，其中 $M1$ 与 $S3^*$ 在长、宽和通道数上相等。最后将 $M1$ 与 $S3^*$ 按照对应元素相加融合后得到 $M2$。同理计算得到融合特征 $M3$ 和最终输出 out。整个计算过程可以表示为式(6)，式中 Concat 表示特征图在通道上的拼接计算，F_{conv} 表示卷积核为 3×3，步距为 2 的卷积计算。

$$\begin{cases} M1 = F_{conv}(S2^*),\ M1 \in \mathbb{R}^{28\times28\times512} \\ M2 = F_{conv}(\text{Concat}(S3^*, M1)),\ M2 \in \mathbb{R}^{14\times14\times1024} \\ M3 = F_{conv}(\text{Concat}(S4^*, M2)),\ M3 \in \mathbb{R}^{7\times7\times2048} \\ out = \text{Concat}(F_{conv}(M3), S5),\ out \in \mathbb{R}^{7\times7\times2048} \end{cases} \tag{16}$$

模型损失计算由分类损失和回归损失两部分构成。分类损失计算流程如下：首先将特征融合输出特征图输入到 3 个全连接中分别计算得到 $M4$、$M5$ 和 $M6 \in R^{1\times66}$，计算过程可以表示为式(7)，式中 FC 表示全连接计算：

$$Mi = FC(out),\ i \in [4,5,6],\ Mi \in \mathbb{R}^{1\times66} \tag{7}$$

其次用 Softmax 函数分别计算 3 个向量类别概率值，Softmax 计算过程如式(8)所示。三个类别中最大的概率值对应 Pitch、Roll 和 Yaw 3 个角度的预测分类结果。最后用估计的分类概率值和真实类别标签计算交叉熵损失，分类损失计算过程如式(9)所示。

$$\hat{\boldsymbol{y}}_n = \frac{e^{z_n}}{\sum_{j=1}^{C} e^{z_j}},\ n \in [0,66) \tag{8}$$

$$L_{\text{class}} = -\frac{1}{BN}\sum_{b=1}^{B}\left(\sum_{n=1}^{N}\left[\boldsymbol{y}_n \log\hat{\boldsymbol{y}}_n + (1-\boldsymbol{y}_n)\log(1-\hat{\boldsymbol{y}}_n)\right]\right)_b \tag{9}$$

式中，Z_n 表示全连接网络输出向量第 n 个值；$\hat{\boldsymbol{y}}_n$ 表示 Softmax 计算出第 n 个类别的概率的预测值；$\boldsymbol{y}_n$ 表示第 n 个真实类别标签；N 表示类别个数；B 表示批训练数量；L_{class} 为分类损失。

回归损失计算是通过对 $\hat{\boldsymbol{y}}$ 积分运算得到估计的头部姿态估计结果与真实标签值计算平方差值损失，计算过程如式(10)所示。

$$L_{\text{reg}} = \frac{1}{B}\sum_{b=1}^{B}\left(\sum_{n=1}^{N}(n\times\hat{\boldsymbol{y}}_n) - \boldsymbol{T}_b\right)_b^2,\ N = 66 \tag{10}$$

式中，B 表示批训练数据数量，$\boldsymbol{T}_b$ 表示批训练集中第 b 个头部姿态角标签值，L_{reg} 表示回归损失。

将交叉熵损失与乘以权值系数的回归损失相加得到总损失，其计算过程如式(11)式所示，其中 α 为权重系数，由文献[13]可知常用的取值为 1 或 2，L_{sum} 为总损失。

$$L_{\text{sum}} = L_{\text{class}} + \alpha L_{\text{reg}} \tag{11}$$

3 实验与分析

3.1 数据集的选取和数据增强

在本文实验中，模型训练和验证都需要接近真实场景且精确标注头部姿态信息数据集。这种场景数据能够真实反映模型的鲁棒性和性能，使得训练的模型适用于真实场景且方便和别的算法模型在性能上进行对比。根据以上需求选取数据集 AFLW2000[15] 和 300W _ LP[16] 用于分析和验证。其中 AFLW2000 数据集来源于 AFLW 数据集前 2 000 张图片，且这些数据使用 3D 模型拟合人脸的方法精确标注 68 点 3D 人脸标志点和头部姿态信息；300W _ LP 包含主流的 2D 人脸关键点标注的数据(AFW、LFPW、HELEN、IBUG 和 XM2VTS)，通过 3D 六自由度人脸模型和 2D 关键点建立映射关系并在偏航角上对头部进行偏角变换的方法扩增数据集，一共得到 61 225 张图片。2 个数据集的图片和真实标签可视化如图 5 所示，图 5(a)表示 AFLW2000 数据集和标签值可视化，图 5(b)表示 300W _ LP 数据集和标签值可视化。根据右手坐标系原则面部朝向为 Z 轴，X、Y 和 Z 坐标轴分别指向右、下和前。

为了减小模型对强烈光照变化、图像形变和随机遮挡带来的影响，在训练过程中通常对数据进行预处理，即人为增加图像复杂度。这一方面使得在少量数据集的情况下得到大规模数据集训练的效果，另一方面可以减少模型的过拟合，提高模型泛化能力。本文在数据数据预处理方面加入的策略如图 5所示。

3.2 实验环境搭建与模型训练设置

本文实验是在 Ubuntu 18.04LTS 操作系统上搭建实验环境，深度学习框架为 Pytorch V1.6.0 并安装 OpenCV-Python V4.5.4、Numpy V1.21.4 等数据处理软件包，加速并行计算采用 CUDA 9.1 架构平台和 CuDNN 7.5 软件包。硬件部分采用具有 CPU Intel Core™ 2.9 GHz i5 9400F(16 GB 运行内存)和 GPU NVIDIA GeForce 1660 (6 GB 显存)的计算机。

特征网络参数初始化采用 ImageNet[17] 预训练模型初始化，新增的网络层采用文献[18]提出的初始化方式设置权重。数据集训练 30 轮次(epoch)，训练时将 300W _ LP 分为训练集和测试集，比例为 9∶1。AFLW2000 作为测试数据集。训练初始学习率设为 0.001，8 ~ 17 轮(epoch)和 18 ~ 30 轮(epoch)的学习率分别减小 0.100 和 0.001 倍，动量因子设置为 0.9，热力图可视化方法使用文献[19]提出的 Grad-CAM 方法。

在模型精度评价指标上，采用平均误差绝对值 *Error* 对模型预测精度进行评估。其值越小表示预测结果越精准，计算过程如式(12)所示。

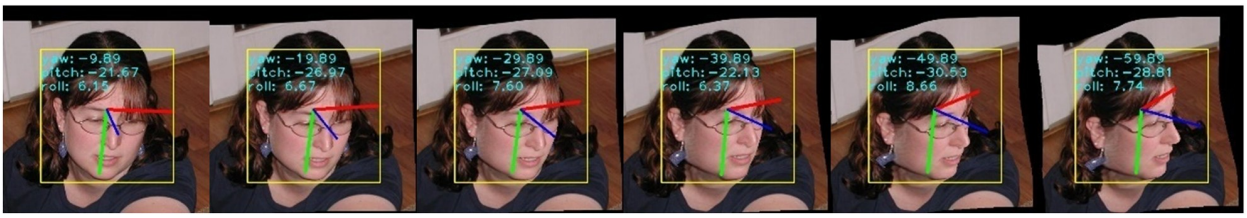
$$Error = \frac{1}{N}\sum_{n=1}^{N} | P_n - P_n' | \tag{12}$$

式中，P_n 第 n 张图片真实值，P_n' 第 n 张图片的预测值，N 为预测图片数量。

3.3 模型改进前后训练损失和验证集误差曲线对比

为了探究本文优化后的模型相对于 HopeNet 在训练损失和平均绝对值误差上的优劣，绘制了改进

(a) AFLW2000 数据集示例



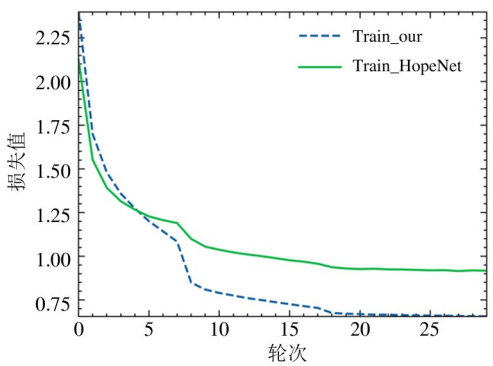
(b) 300W _ LP 数据集示例

(c) 随机擦除　　(d) 灰度变换

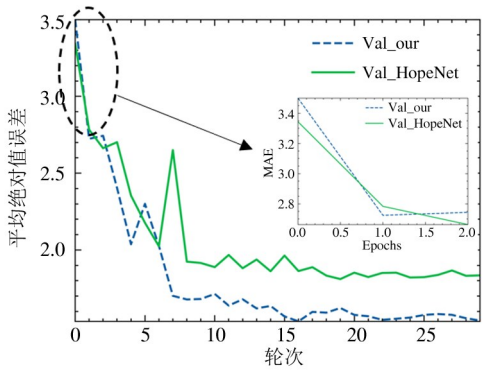
(e) 亮度变换　　(f) 对比度变换

图 5　AFLW2000 和 300W _ LP 数据集可视化与数据增强

前后的模型在 300W _ LP 数据集上训练后得到损失值和平均绝对值误差(mean absolute error,MAE)随着训练轮数增加的变化曲线,结果如图 6 所示。从图 6(a)中可以看到,从第 5 轮训练开始,HopeNet 损失值下降相对缓慢,而本文方法的损失值继续快速下降;在训练的第24轮以后2个方法的损失曲线都



(a) 训练损失曲线



(b) 平均绝对值误差曲线

图 6　本文方法和 HopeNet 的训练损失与误差曲线

趋于平缓,说明 2 个模型训练没有出现过拟合且达到最优。观察图 6(b)验证集上的平均绝对值误差曲线可知,从开始训练到第 10 轮期间优化前后模型曲线都存在一定波动,但是本文方法波动更小。本文优化模型在第 20 轮以后平均绝对值误差趋于平稳。由此可见,本文优化方法训练得到的模型损失值更小,训练过程中波动更小且平均绝对值误差更小。

3.4　主流方法对比

为了验证本文方法与主流方法相比在预测误差上的优越性而设置了对比实验。其中主流方法主要包括基于先验信息和无需先验信息的头部姿态估计 2 个大类。所述方法在 AFLW2000 数据集上预测结果如表 1 所示。虽然需要先验信息的头部姿态方法(Landmarks、3DDFA、FAN(12point)和 Dlib(68point))的精度在不断提高,但是精度略低于无需先验信息方法(HopeNet 和 FSA-Net)。本文方法和无需先验信息方法进行对比得出,前者在 Yaw、Pitch 和 Roll 3 个角度估计上精度有一定的优势,与现今精度最好的 FSA-Net 相比其平均误差减小了 0.87 °,与 HopeNet 相比其平均误差减小1.90 °。通过上述对比分析,可以看出本文方法在头部姿态估计精度上有一定的优越性。

表 1　主流的头部姿态估计方法在 AFLW2000 数据集上预测误差对比

模型	Yaw/°	Pitch/°	Roll/°	MAE/°
Dlib(68point)[20]	23.10	13.60	10.50	15.80
Landmarks[21]	5.90	11.80	8.20	8.65
FAN(12point)[22]	6.35	12.20	8.70	9.10
3DDFA[16]	5.40	8.53	8.25	7.39
Hopenet(α=1)[13]	6.90	6.60	5.60	6.40
Hopenet(α=2)[13]	6.50	6.50	5.40	6.10
FSA-Net[23]	5.10	4.50	4.78	5.07
本文方法	**5.00**	**3.80**	**3.80**	**4.20**

3.5　消融实验

为了验证在 HopeNet 的基础上单独增加改进特征激励模块(SE)或特征融合模块(Fusion)和同时增加 2 个模块对整体模型精度的影响,设置了消融实验。在参数量、推理时间和平均误差上进行对比,实验结果如表 2 所示,表中的模型是在 300W_LP 数据集上训练得到,验证集误差是在 AFLW2000 数据集上验证得到。由表 2 可知,虽然 SE 模块和 Fusion 模块分别增加了 2.53×10^6 和 14.42×10^6 的参数量和少量推理时间,但是 2 个模块可以分别实现 0.53 °和 1.07 °的平均误差减小。将 2 个模块同时引入模型中可以减小 1.90 °的平均误差,且优于任何一个模块的单独使用。由此得出结论,添加本文提出的特征融合模块或改进的注意力机制模块对精度提升有促进作用,且叠加使用时精度提升优于任何一个模块的单独使用。

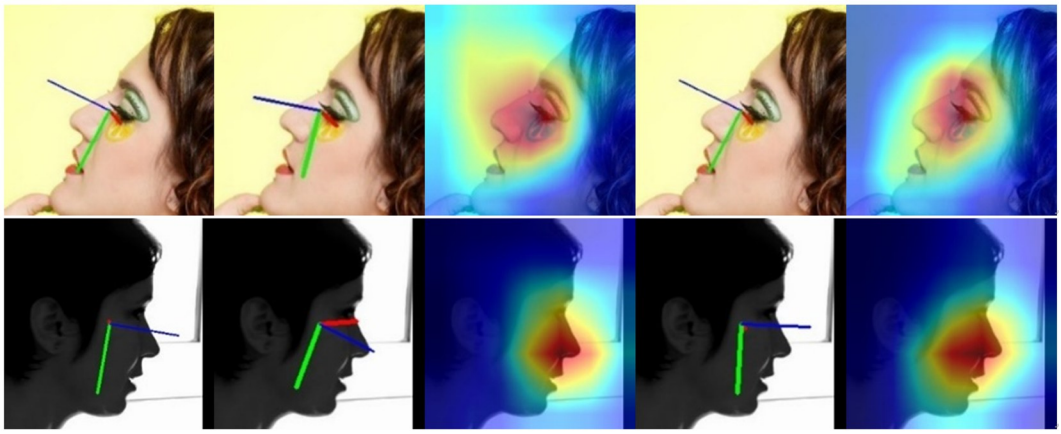
表 2　增加压缩激励和融合模块后对模型性能影响

模型	参数量/$\times10^6$	时间/s	MAE/°	ΔMAE	精度提升/%
HopeNet(α=2)	23.92	0.011	6.10	0.00	0.00
HopeNet(α=2)+SE	26.45	0.014	5.57	+0.53	8.68
HopeNet(α=2)+Fusion	38.34	0.023	5.03	+1.07	17.54
HopeNet(α=2)+SE+Fusion	40.87	0.028	4.20	+1.90	31.15

3.6　网络推理的热力图分析与预测结果可视化对比

探究本文提出方法对精度提升的内在原因,将本文方法和 HopeNet 的预测结果和卷积响应热力图可视化分析,结果如图 7 所示。由图 7(b)列和图 7(d)列可以看到,本文方法在大角度和大尺度图像上预测结果和标签值更加接近。由图 7(c)列和图 7(e)发现,本文方法特征响应更加集中于面部信息,而 HopeNet 方法的卷积热力响应图更加发散,使得背景特征信息更多贡献于预测结果,导致精度较差。

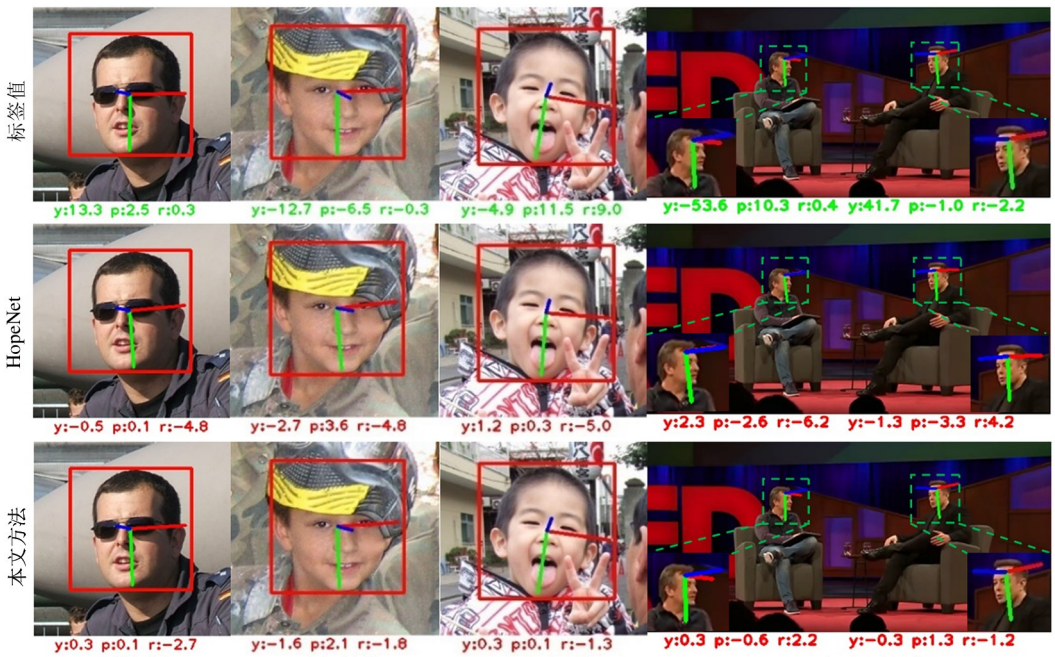
上述讨论验证了本文改进方法使得模型更加关注于面部信息,减少背景信息的影响,从而提升了模型预测精度。



(a) 标签值　(b) HopeNet结果　(c) HopeNet卷积　(d) 本文方法结果　(e) 本文方法卷积响应热力图响应热力图

图 7　本文方法与 HopNet 的预测结果和卷积响应热力对比

为了对比改进前后模型在实际场景中主观可视化效果和预测姿态角误差大小,将 HopeNet 和本文优化方法分别在真实场景图像上预测的结果进行对比和分析。改进前后模型预测误差与标签值可视化如图 8 所示,“标签值”所在行对应图片中字母 y 表示 Yaw、p 表示 Pitch、r 表示 Roll;“HopeNet”和“本文方法”所在行对应图片中字母后的数值表示当前图中预测姿态角与标签姿态角误差值(误差 = 预测值 - 真实值);头部姿态可视化表示与坐标轴朝向已在 3.1 节中说明。由于模型的输入为头部图像,所以借助人脸检测算法得到人脸框,然后适当放大和平移人脸框,裁剪出头部图像送入头部姿态估计模型。由于人脸识别不是本文讨论重点所以人脸检测采用 MTCNN[24] 方法实现。一共测试了 4 种场景图片,图 8(a)戴墨镜场景中由于面部特征信息被遮挡从而影响预测精度,前者 Roll 方向误差比本文方法大 2.1°,本文方法与标签值接近;图 8(b)头部遮挡场景中由于头部特征被遮挡对 HopeNet 方法精度影响较大,但是本文模型在应对这种场景上有一定的优势,在Yaw、Pitch和Roll角度误差上比前者分



(a) 戴墨镜　(b) 头部遮挡　(c) 特殊面部　(d) 小尺度场景

图 8　实际场景头部姿态预测对比

别小 1.1 °、1.5 °和 3 °;图 8(c)特殊面部场景中由于预测的面部图像发生形变而影响特征提取,本文方法在 Yaw 和 Pitch 方向误差较小,分别为 0.3 和 0.1,与真实值比较接近,HopeNet 方法受到图中手部和面部形变的影响在 Roll 旋转方向误差较大,达到 5.0 °,而本文方法误差仅有 1.2 °;图 8(d)小尺度场景中同时预测 2 个头部姿态,由于面部和头部关键信息减少导致特征提取困难,但是本文方法在 2 个头部姿态的 Roll 方向误差优于 HopeNet,前者比后者在误差上分别减少 4.0 °和 3.0 °,在另外 2 个旋转角误差上本文方法与标签值更为接近。通过上述分析可以得出,本文改进的模型,即在 HopeNet 基础上增加改进特征压缩激励和本文提出的特征融合机制,能够应对面部遮挡和面部形变特征提取困难而导致精度下降的问题,预测出的旋转角误差更小,应对小尺度场景特征信息大量减少的场景也有一定的优越性,更符合实际场景需求。

4 结 论

为了提升基于无需先验信息的头部姿态估计在复杂背景图像和多尺度图像场景下的精度,本文提出了一种基于改进 HopeNet 的头部姿态估计方法。在残差网络结构中增加特征压缩激励模块,使得网络具有自适应学习特征层重要程度权重,让模型更加关注目标图像区域。通过添加特征融合模块使得模型能够同时利用网络的深层特征信息与浅层特征信息,使得网络能够充分利用特征信息。实验结果表明,从在 AFLW2000 数据集上的预测结果来看,与 HopeNet 相比本文方法在平均误差上降低了 1.90 °,与主流算法相比本文方法在精度上有一定的优越性。通过卷积响应热力图可视化分析可知,改进后的网络更加关注面部特征,减少了背景特征对模型精度的影响。通过在实际场景图像上的预测结果可视化分析得出,对于面部遮挡和小尺度图像场景,本文方法在预测精度上有一定的优势。今后,如何在减少计算量的同时保持精度稳定使得算法适合在小型嵌入式设备运行将是下一步的研究重点。

参考文献

[1] BEVILACQUA V, ANDRIANI F, MASTRONARDI G. 3D head pose normalization with face geometry analysis, genetic algorithms and PCA[J]. Journal of Circuits, Systems, and Computers, 2009,18(8):1425-1439.

[2] MURPHY-CHUTORIAN E, DOSHI A, TRIVEDI M M. Head pose estimation for driver assistance systems: a robust algorithm and experimental evaluation[C]//2007 IEEE Intelligent Transportation Systems Conference. Las Vegas, USA: IEEE, 2007:709-714.

[3] CORDEA M D, PETRIU D C, PETRIU E M, et al. 3D head pose recovery for interactive virtual reality avatars [J]. IEEE Transactions on Instrumentation and Measurement, 2002,51(4):640-644.

[4] 齐永锋, 马中玉. 基于深度残差网络的多损失头部姿态估计[J]. 计算机工程, 2020,46(12):247-253.

[5] 钟俊宇, 邱健, 韩鹏, 等. 基于结构光三维重建的头部姿态估计算法[J]. 激光与光电子学进展, 2020,57(18):112-119.

[6] 梁令羽, 张天天, 何为. 多尺度卷积神经网络的头部姿态估计[J]. 激光与光电子学进展, 2019,56(13):131003.

[7] DROUARD V, HORAUD R, DELEFORGE A, et al. Robust head-pose estimation based on partially-latent mixture of linear regressions[J]. IEEE Transactions on Image Processing, 2017,26(3):1428-1440.

[8] GENG X, XIA Y. Head pose estimation based on multivariate label distribution[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014:1837-1842.

[9] HUANG C, DING X, FANG C. Head pose estimation based on random forests for multiclass classification[C]//The 20th International Conference on Pattern Recognition. Istanbul, Turkey:IEEE, 2010:934-937.

[10] PATACCHIOLA M, CANGELOSI A. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods [J]. Pattern Recognition, 2017,71:132-143.

[11] ZHOU Y, GREGSON J. WHENet: real-time fine-grained estimation for wide range head pose[EB/OL]. (2020-05-20)[2023-02-25]. https://arxiv.org/pdf/2005.10353.pdf.

[12] BERRAL-SOLER R, MADRID-CUEVAS F J, MUNOZ-SALINAS R, et al. RealHePoNet: a robust single-stage ConvNet for head pose estimation in the wild[J]. Neural Computing and Applications, 2021,33(13):7673-7689.

[13] RUIZ N, CHONG E, REHG J M. Fine-grained head pose estimation without keypoints[C]//IEEE Conference

on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA : IEEE, 2018:2074-2083.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7132-7141.

[15] YIN X, YU X, SOHN K, et al. Towards large-pose face frontalization in the wild[C]// IEEE International Conference on Computer Vision. Venice, Italy:IEEE, 2017: 3990-3999.

[16] ZHU X, LEI Z, LIU X, et al. Face alignment across large poses: a 3D solution[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:146-155.

[17] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015,115(3):211-252.

[18] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification[C]// IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1026-1034.

[19] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]// IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 618-626.

[20] KAZEMI V, SULLIVAN J. One millisecond face alignment with an ensemble of regression trees[C]// IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA:IEEE, 2014:1867-1874.

[21] RUIZ N, CHONG E, REHG J M. Fine-grained head pose estimation without keypoints[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE, 2018: 2155-2164.

[22] BULAT A, TZIMIROPOULOS G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3D facial landmarks)[C]// IEEE International Conference on Computer Vision. Venice, Italy:IEEE, 2017:1021-1030.

[23] YANG T Y, HUANG Y H, LIN Y Y, et al. SSR-Net: a compact soft stagewise regression network for age estimation[C]// The 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press, 2018:1078-1084.

[24] XIANG J, ZHU G. Joint face detection and facial expression recognition with MTCNN[C]// The 4th International Conference on Information Science and Control Engineering. Changsha, China:IEEE, 2017:424-427.

Head pose estimation method based on improved HopeNet

ZHANG Liguo, HU Li

(Measurement Technology and Instrumentation Key Laboratory, Yanshan University, Qinhuangdao 066004)

Abstract

Aiming at the poor accuracy of the head pose estimation algorithm based on no prior knowledge in complex background images and multi-scale image scenes, a head pose estimation method based on improved HopeNet is proposed. Firstly, the feature fusion structure is added to the backbone network structure to make the model make full use of the deep and shallow feature information of the network and improve the feature analysis power of the model. Then feature squeeze and excitation module is added to the residual structure of the backbone network, so that the network can adaptively learn the weight information of different feature layers and the model can pay more attention to the target information. Experimental results show that compared with HopeNet, the accuracy of the proposed method on AFLW2000 dataset is improved by 31.15%, and the average error is reduced to 4.20°. Meanwhile, the proposed method has good robustness in complex background image scenes.

Key words: head pose estimation, HopeNet, characteristics of the fusion, characteristic compression and excitation, adaptive learning