基于增强头部姿态估计的人脸表情识别模型

崔景春 王 静
(南京工业大学计算机科学与技术学院 南京 211816)

摘 要 针对现有表情识别算法未考虑头部姿态及不能使用高像素图像的问题,提出一种基于随机森林算法的头部姿态估计(RF-HPE)网络与卷积神经网络相结合的模型。首先对输入图像作强度归一化,然后利用RF-HPE确定脸部标志关键点,从而确定脸部标志的位置,最后使用卷积神经网络提取特征并训练模型。该模型降低了光线强度对识别结果的影响,并且在不牺牲算法效率的情况下提高了训练精度。实验结果表明,所提出的改进模型的学习能力相比其他同类模型有较大优势,分类精度也显著提高。

关键词 强度归一化,随机森林算法,头部姿态估计,表情识别
中图法分类号 TP183,TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.049

Face Expression Recognition Model Based on Enhanced Head Pose Estimation

CUI Jing-chun WANG Jing
(School of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

Abstract Aiming at the problem that the existing expression recognition algorithm does not consider the head pose and cannot use the high-pixel picture, this paper proposed a model based on the random forest algorithm-head pose estimation (RF-HPE) network combining a convolutional neural network. First, the input image is normalized by intensity. Then the key points of the face marker are determined by using RF-HPE to determine the position of the face marker. Finally, a convolutional neural network is used to extract features and train model. This model reduces the influence of light intensity on the recognition result and improves the training accuracy without sacrificing the efficiency of the algorithm. Experimental results show that the improved model has greater advantages than other similar models, and the classification accuracy is also significantly improved.

Keywords Intensity normalization, Random forest algorithm, Head pose estimation, Expression recognition

1 引言

人脸识别是机器学习中非常经典的技术,对它的研究最早出现于20世纪60年代,经过多年的发展,其已被应用于多个方面,如视觉监控、身份认证、反恐侦查等[1]。表情识别是人脸识别的应用之一,随着互联网的发展,越来越多的用户在社交网络上分享自己的照片。为了能更好地利用这些资源,表情识别越来越受到研究人员的关注[2]。表情识别最早起源于20世纪70年代,但早期主要集中在心理学和生物学方面。随着计算机技术的发展以及机器学习技术的出现,从人脸图像中学习人的情感成为可能[3-4]。

人脸表情识别主要分为3个步骤:首先是人脸检测与定位,然后对脸部标志特征进行识别和提取,最后对提取出的特征进行学习分类[5]。对于最后的分类大多研究选用统计学方法,选用何种方法依具体情况而定,在此不予讨论。近10年,人脸检测与定位来取得了长足的发展和一定的成功,Girshick等使用卷积神经网络(Convolutional Neural Network,CNN)进行人脸检测[6],在图像较小时能获得较好的效果,但在处理分辨率较高的图像时计算量大大增加,精度也会有所降低。Zhou等预先将大图片分割成不同的部分,对每部分使用CNN进行学习[7],能取得较好的效果,但该方法仍不能在高分辨率图像上取得良好的效果。表情特征识别和提取是表情识别的关键步骤,Kotsia等提出使用主成分分析法(PCA)进行特征提取[8],以排除特征中的冗余信息,然而该方法对存储空间有很大的要求,而且未考虑高阶统计信息,在特征提取时可能丢失重要信息。Pietikinen等提出使用局部二值模式(LBP)来提取特征[9],通过图像像素间的关系形成一个二进制编码,最后采用多区域直方图作为特征描述。该算法的复杂度低、消耗内存小,然而无法区分领域像素点与中心像素点相等时的情况。现有的方法多是对人脸图像直接进行定位操作,没有考虑图像的亮度、角度因素。

本文使用随机森林(Random Forest,RF)算法进行头部

到稿日期:2018-05-21 返修日期:2018-09-27 本文受南京工业大学引进人才启动基金项目(39809110)资助。
崔景春(1993-),男,硕士生,主要研究方向为机器学习与深度学习;王 静(1982-),女,副研究员,主要研究方向为无线传感器网络技术,E-mail:wj_cec@126.com(通信作者)。

姿态估计(Head Pose Estimation，HPE)，提出一种基于随机森林算法的头部姿态估计网络(Random Forest-Head Pose Estimation，RF-HPE)来进行人脸检测与标志点的定位。与已有算法相比，本文的模型在整个网络之前加入灰度归一化的操作，可以降低图像亮度等环境因素的影响；使用了随机森林算法，在相同的准确率下，计算量显著减少；同时，引入了头部姿态估计网络，使得本网络不仅可以检测图像中的人脸，还可以考虑到头部的旋转、俯仰等信息。利用 RF-HPE 的输出来辅助表情特征的提取，可以提高有效特征的提取率。

2　相关介绍

2.1　卷积神经网络

卷积神经网络是一种深度学习网络，最早由 Lecun 在 1998 年提出[10]，通常包括特征提取层、下采样层以及一个顶层分类器。CNN 最早应用在手写字体识别任务中，随后在图像识别、语言识别等领域得到了广泛应用。CNN 通常采用局部连接和权值共享的方式，一方面减少了网络中需要计算的权值数量，另一方面降低了网络发生过拟合的可能，这一特性使得 CNN 非常适用于处理图像这种维度较高的问题。

2.2　图像强度

图像强度(image intensity)表示单通道图像像素的强度(值的大小)。在灰度图像中，图像强度指的是图像的灰度。在 RGB 颜色空间中，图像强度可以理解为 R，G，B 3 个通道之一的像素灰度值，也就是 RGB 颜色有 3 个图像强度。图像强度最直观的表现就是图像的亮度和对比度，对于同一个人的表情图片，由于光线强度不同，其图像强度也不同，转化为特征向量时也会有所不同，因此增加了特征向量的变化，这种变化增加了分类器解决问题的复杂性。为了提高图像识别的准确性，需要采取必要的措施来降低该因素的影响[11]。

2.3　随机森林算法

随机森林(RF)算法是一种较灵敏的机器学习算法，其关键是随机森林的建立，该森林由许多棵相互之间没有关联的决策树组成。在森林建立之后，每当有一个新的样本输入森林时，就使其中的所有决策树都判断此样本属于哪个类别，然后统计数量最多的类别，将此样本判定为该类别。它既适用于分类问题，也适用于回归问题[12]。使用机器学习处理图像问题时，转化出的特征向量维度很高，需要进行降维操作，否则计算量太大。而采用随机森林算法，由于引入了随机性，使得随机森林不易陷入过拟合且具有良好的抗噪效果，能够有效应用于大数据集上，可以处理特征极多的高维数据且不需要降维，而且无需对数据集进行规范化处理，因此非常适合处理维度较高的图像问题。

3　基于增强头部姿态估计的 CNN 模型

本节详细阐述了本文提出的模型，图 1 为模型的结构图。可以看出，模型由头部姿态估计网络和表情识别网络组成，表情识别网络包含特征提取网络。首先通过强度归一化将原始人脸图像归一化为灰度图，以降低光亮和对比度对后续步骤的影响。然后将生成的灰度图分别送入头部姿态估计网络和表情识别网络。头部姿态估计网络基于随机森林算法训练而来，其可估计人脸图像中的头部姿态信息。表情识别网络首先从灰度图中提取出人脸的轮廓模型，然后将头部姿态信息和脸部轮廓特征同时传入局部特征提取层，将得到的局部特征进行降维后送入顶层的 softmax 分类器。

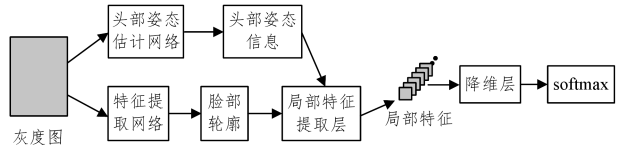


图 1　基于头部姿态估计的卷积神经网络模型

Fig. 1　Convolutional neural network model based on head pose estimation

3.1　强度归一化

现有的表情识别模型大多使用原始图像作为输入，没有对原图进行必要的预处理来降低环境因素的影响。为了降低表情识别过程中图像强度对识别准确率的影响，本文采用强度归一化方法将图像都转化为灰度图，以降低对比度和亮度对特征提取的影响。强度归一化分为两个过程：首先减去像素的局部归一化平均值；其次，采用分裂的局部对比度归一化。在第一步中，每个像素的值从其相邻的高斯加权平均值中减去。在第二步中，每个像素值除以其邻域的标准差。这两个过程的邻域使用 7×7 像素的内核(经验选择)。图 2 举例说明了该过程。

图 2　强度归一化过程

Fig. 2　Process of intensity normalization

在强度归一化过程中计算每个新的像素值的公式如下：

$$x'=\frac{x-\mu_{nhgx}}{\sigma_{nhgx}} \tag{1}$$

其中，x'是新的像素值，x 是原始像素值，μ_{nhgx}是 x 相邻位置像素的高斯加权平均值，并且 σ_{nhgx}是 x 相邻位置像素的标准偏差。

3.2　RF-HPE

表情识别问题的关键是要提取出人脸的几个关键部位来进行训练，想要提取出准确的局部位置，需要找出描述局部位置的关键点，而头部的姿态信息可以用来确定这些关键点。为了获得更准确的头部姿态信息，从而提高局部特征提取的准确性，本文采用 RF 算法来确定头部姿态并将其应用于后续的表情识别任务。

可以用一组 $P=(\phi_x,\phi_y,\phi_z,t_x,t_y,t_z)$的六维向量来刻画头部姿态，该向量由轴角与鼻尖的位置两个参数来确定，即 (ϕ_x,ϕ_y,ϕ_z)表示头部偏转与俯仰的角度，(t_x,t_y,t_z)是鼻尖的位置坐标。头部朝向的轴角作如下说明：$(\phi_x,\phi_y,\phi_z)\leftrightarrow((e_x,$

$e_y,e_z),\theta)\leftrightarrow(\theta\cdot e_x,\theta\cdot e_y,\theta\cdot e_z)$，其中$(e_x,e_y,e_z)$是一个单位旋转轴向量，$\theta$是相应的旋转角。

头部姿态估计网络的输入是一组深度图片的集合，首先将此图片集合转化为头部点云集合。当随机森林中的任意决策树需要训练时，都是在此点云集合中随机抽取一定比例的样本。对于任意的头部点云，用包围盒在其中随机地选取一定数量的点云块，包围盒的大小一般设为人脸的 1/4 左右。因此每一个训练样本都有自己对应的一组点云块，训练一棵树所需要的样本$\{L_i,\Theta_i\}_{i=1}^{N}$就是对应点云块的集合，其中$\Theta_i=\{\Phi_{ix},\Phi_{iy},\Phi_{iz},\Delta t_{ix},\Delta t_{iy},\Delta t_{iz}\}$是每个点云块带有的参数信息，$\{\phi_{ix},\phi_{iy},\phi_{iz}\}$是头部姿态真实值(groundtruth)向量，表示第 i 个点云块所处点云的实际头部方向向量，$\{\Delta t_{ix},\Delta t_{iy},\Delta t_{iz}\}$是第 i 个点云块的中心位置到其所处点云的实际鼻尖位置的差值向量。

以训练一棵树的过程为例来说明随机森林的训练过程。一个非叶结点(non-leaf node)处二分的分裂准则为：

$$bt_{L_{i,\tau}}=D(l_i^1)-D(l_i^2)>\tau \tag{2}$$

其中，$D(l)$表示点云块 l 包含的所有点云的深度平均值，$\{l_i^1,l_i^2\}$是包含于点云块 L_i 中的特征点云对，τ 是随机选定的阈值。分裂准则越准确，分类到两个叶子结点(children-node)中的样本集的差异就越大，且同一叶子结点的样本集差异也尽可能小。因此，通过最小化式(3)来优化二分准则。

$$E=\frac{|L|}{|L|+|R|}\cdot(\sigma_L^{\phi}+\sigma_L^{\Delta t})+\frac{|R|}{|L|+|R|}\cdot(\sigma_R^{\phi}+\sigma_R^{\Delta t}) \tag{3}$$

其中，$|L|$和$|R|$表示在分裂准则 $bt_{L_{i,\tau}}$ 的约束下训练样本集分别落入左右两个叶子结点中的数量，$\sigma_{\Delta t}$表示落入一个叶子结点的全部样本集包含的点云块中心到鼻尖位置的位移矢量的协方差矩阵的迹，σ_{ϕ}表示落入该叶子结点中的所有训练样本集的头部姿态方向的协方差矩阵的迹。

确定了一个非叶结点的最优二分准则后，根据式(2)的条件，将落入该结点的全部训练样本集分类到其两个叶子结点中。在一棵树的所有非叶结点上，不断迭代该训练过程，直至该树的深度达到设定好的最大深度，或者结点的样本数量达到设定的最小值时训练停止，此结点即为树的一个叶子结点(leaf node)。训练结束即代表着随机森林中的一棵树构建完成，随机森林就是由许多棵这样的树组成。每个叶子结点都保存着许多包含实际值信息标签的训练集，表示为形如$\{\Theta_i\}=\{\{\Phi_x,\Phi_y,\Phi_z,\Delta t_x,\Delta t_y,\Delta t_z\}_i\}$的向量。此时，能够将这些向量的平均值和协方差矩阵(表示为 $\boldsymbol{\Sigma}$)计算出来并保存在原来的结点中。

经过上述训练过程，对于每一张包含人脸的深度图像，通过选取的点云块都可以训练出其头部姿态估计值，而一张脸部图像会进行多次点云块的选取训练，因此最终训练得到的是一组包含头部姿态估计值的向量集合。然后对集合中的全部向量进行聚类，本文采用的聚类算法是均值漂移算法(meanshift)。最后计算聚类后的平均向量，得到最终需要的头部姿态估计值。图 3 是基于随机森林算法的头部姿态估计过程的说明。

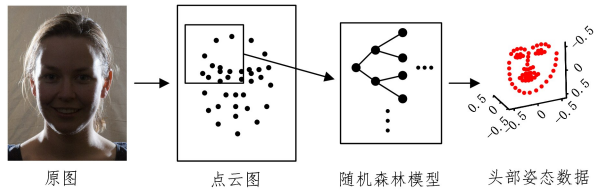


图 3 基于 RF 的头部姿态估计过程

Fig. 3 Process of head pose estimation based on RF

3.3 基于 RF-HPE 的 CNN 模型

模型由两个网络组成，分别是头部姿态估计网络和表情识别网络。头部姿态估计网络的训练过程前面已详细说明，通过头部姿态估计网络能够得到一组包含头部姿态估计信息的向量，即 $P=(\phi_x,\phi_y,\phi_z,t_x,t_y,t_z)$，由该向量可以得到两个信息，分别是脸部中心(即鼻尖的位置)和脸部朝向角度，这两个信息可用于后续脸部标志点的确定。

表情识别网络的本质是一个卷积神经网络(CNN)。卷积运算提取脸部特征是一个由低级到高级的过程，第一层提取到的是脸部边缘的线，在随后的层中提取越来越抽象的特征。为了提取出图像中的脸部轮廓，本文模型中的特征提取网络包含 3 个卷积层，其卷积核大小分别为 7×7，3×3，3×3。为了保留完整的局部特征，设置卷积的步长为 1，并设置零填充大小，使得图像的宽度和高度保持不变。卷积运算如式(4)所示：

$$x_j^i=f\Big(\sum_{i\in M_j}x_i^{i-1}\times k_{ij}^i+b_j^i\Big) \tag{4}$$

其中，M_j 表示输入的 maps 集合，k 为权值，b 为偏差向量，$f(\ )$为非线性激活函数。本文采用 ReLu 作为激活函数，对于非线性函数而言，ReLu 可以避免消失梯度问题，使得模型的收敛速度维持在稳定状态，并且可显著减少收敛所需的迭代次数，其表示形式为：

$$ReLu=\begin{cases}x, & x>0\\ 0, & x\leqslant 0\end{cases} \tag{5}$$

通过特征提取网络，可以确定面部图像中的脸部轮廓位置，然后结合从头部姿态估计网络中获得的信息来确定标志点的数量与位置，用这些标志点可以确定脸部图像中标志位置的初始估计。初始位置估计如式(6)所示：

$$x^i=\begin{cases}sR(\theta)PX_{3D}+T, & i=0\\ U(s,\sigma_s)R(U(\theta,\sigma_\theta))PX_{3D}+U(T,\sigma_T), & i>0\end{cases} \tag{6}$$

其中，s,θ 和 T 是由头部姿态估计网络估计的比例、旋转和平移参数；R 是旋转矩阵；P 是正交投影矩阵；$U(a,b)$是均匀分布的采样值，代表 $unif(a-b,a+b)$；i 是初始形状索引；σ_s，σ_θ和σ_T表示均匀分布的扰动值；X_{3D}是 3D 标志人脸模板模型。此外，设置以下参数：$\sigma_s=0.05$s，$\sigma_{\theta_x}=10°$，$\sigma_{\theta_y}=5°$，σ_{T_x}，$\sigma_{T_y}=10$ 像素。由于头部姿态估计网络的初始预测可能包含误差，需从均匀分布中获得多个初始形状来增强该估计的准确性，因此通过平均多个初始形状的结果(本文使用 10 个)来获得最终估计。

在获得标志位置的初始估计之后，通过局部特征提取层提取全局脸部特征空间位置上的局部特征作为输入，然后生

成局部特征描述符。局部特征描述符输入具有欧几里得损失函数的线性回归层，回归层要学习的目标值是泛型下降方向图，它是从当前估计的标志位置到实际位置的矢量。回归量相当于学习它的权重 β，如式(7)所示：

$$\begin{aligned}&\| (l_m \circ l_{m-1} \circ \cdots \circ l_{m-n}(x, l_{m-n-1} \circ \cdots \circ l_1(I))) \\ &\| \hat{\beta} - \Delta x_*^k = \min_{\beta_k} \\ &\| (l_m \circ l_{m-1} \circ \cdots \circ l_{m-n}(x, l_{m-n-1} \circ \cdots \circ l_1(I))) \\ &\| \beta_k - \Delta x_*^k\end{aligned} \tag{7}$$

该式等价于使式(8)的成本函数最小化：

$$E = \min_{\beta_k} \sum (CNN(I_i, x_k^i)\beta_k - \Delta x_*^{ki})^2 \tag{8}$$

其中，N 是样本图像 I 的数量；I_m 是指第 m 层，I_{m-n} 表示局部特征提取层；$\Delta x_*^{ki} = x^{i}_* - x_k^i$，$x^{i}_*$ 是第 i 个样本图像中标志的真实位置；x_k^i 是第 k 次迭代的一个估计标志位置；$CNN()$ 是包含一个偏差节点的局部特征描述符；$\hat{\beta}$ 是通用下降方向图，它是回归的。成本函数最小化是所有脸部标志的估计向量和目标向量之间的差异之和的最小化。

本文采用有监督训练模型，每个图像都有其对应的标签。通常将人脸表情划分为愤怒、蔑视、厌恶、恐惧、快乐、悲伤和惊奇 7 类，每个训练图像都属于其中之一。部分数据库分为 6 类，不包含蔑视类表情。为了便于表示，使用英文缩写来代表各表情，分别为 Anger(An)、Contempt(Co)、Disgust(Di)、Fear(Fe)、Happiness(Ha)、Sadness(Sa)、Surprise(Su)。模型使用反向传播算法调整参数，算法描述如算法 1 所示。

算法 1

1. 输入带有标签 L 的原始人脸图像 I，并使用式(1)对其进行强度归一化处理，得到灰度图 G，设置网络迭代次数 N；
2. 将灰度图 G 送入头部姿态估计网络，通过式(2)进行二分分类；
3. 通过式(3)来优化分类，获得头部姿态信息 P；
4. 同时将灰度图 G 送入特征提取网络，通过式(4)进行卷积运算，获得脸部轮廓图像；
5. 根据头部姿态信息，通过式(6)估计脸部标志特征的初始位置；
6. for i in [1,N] do
7. 　局部特征提取层提取局部特征表示 c=[c_1,c_2,…,c_T]，并进行降维操作；
8. 　将降维后的特征输入 softmax 分类器进行分类；
9. 　根据式(8)计算误差函数，使用反向传播算法调整权值；
10. end for
11. 在测试阶段，使用与训练阶段相同的方式输入图像，网络执行前向传播算法，将网络输出与标签值作比较，计算预测误差来评估模型性能。

4　实验结果和分析

本节评估了所提出的模型在 4 个基准数据集上的表情识别的性能，并将其与其他方法进行对比。本文采用的实验方法是交叉验证方法，即一个数据集中一部分数据用来训练，一部分用来测试[13]。

4.1　基准数据集介绍

1)CK＋：CK ＋是面部表情识别的代表性数据库。该数据库由 327 个图像序列组成，其中包含 7 个情感标签(愤怒、蔑视、厌恶、恐惧、快乐、悲伤和惊奇)，有 118 个主题，这些主题按照 ID 从小到大排序分为 10 组。9 个子集用于训练网络，其余子集用于验证。

2)Oulu-CASIA：Oulu-CASIA 包括在正常照明条件下拍摄的 480 个图像序列。每个图像序列都有 6 种情感标签之一：愤怒、厌恶、恐惧、快乐、悲伤或惊奇。

3)MMI：MMI 由 205 个正面脸部图像序列组成，包含 30 个主题。其类似于 Oulu-CASIA 数据库，有 6 种情感标签。

4)FERA：FERA 由 10 个演员的表情组成，分为 7 组，其中 6 组用于测试，训练集包含 155 个图像序列，测试集包含 134 个图像序列。数据库中共有 5 种情绪类别：愤怒、恐惧、幸福、安心和悲伤。其共有约 7000 个图像。

3 个数据库中各表情图像的数量如表 1 所列。

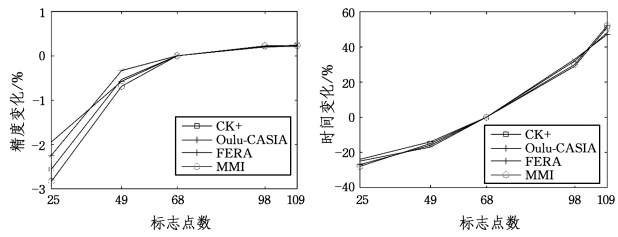
表 1　数据集中各表情图像的数量

Table 1　Number of emoticons in each dataset

	An	Co	Di	Fe	Ha	Sa	Su
CK＋	45	18	59	25	69	28	83
Oulu-CASIA	80	—	80	80	80	80	80
MMI	1959	—	1517	1313	2785	2169	1746
FERA	1681	—	—	1467	1882	2115	—

4.2　参数设置

在模型的训练阶段，参数的设置对网络的性能及训练的效果会产生很大的影响。本模型中，脸部标志点的个数是提取局部特征的关键，为了选取合适的标志点数，将不同的标志点数对准确度以及时间的影响进行比较，以 68 点作为比较的基准，结果如图 4 所示。



(a)　(b)

图 4　标志点数量对模型精确度及时间复杂度的影响

Fig. 4　Effect of landmark point number on model accuracy and time complexity

根据测试结果，随着标志点数的增长，准确度也有所提高，这是因为更多的标志点能够更准确地描述局部特征，从而提高模型的准确率。但是，当标志点数达到 68 以后再进行增加时其准确度的增长幅度很小，而此时所需的训练时间却增长迅速，因此本文设置标志点的数量为 68，以平衡准确率与训练效率。另外，训练批量大小设为 50，CNN 网络的其他参数设置分别如下：dropout 率为 0.5、初始学习率为 1.0、初始衰减因子为 0.9。

4.3　获取头部姿态信息

本模型中的头部姿态估计网络采用基于 K 平均(K-means)算法的无监督模型。首先使用 K-means 算法对训练

集中包含的头部姿态空间进行聚类，本文将头部姿态空间聚为 7 类；然后分别对每一个头部姿态空间训练人脸关键点模型。这种处理很有必要，其原因在于不同头部姿态下采集到的数据在人脸的相同区域的分布存在差别，如不进行区分会极大地增加训练模型的大小，导致测试阶段测试时间的延长。具体训练过程已在 3.2 节做了详细说明。

4.4 对比实验

4.4.1 分表情精确度对比

如果模型对一个表情识别有误，那么就是将原表情误分类为其他表情，为了探究其中的关系，在各个数据集下进行分表情的精确度对比，结果采用矩阵的方式呈现，对角线代表识别准确的比例，列项代表实际表情，行项表示模型分类的结果分布，这样可以直观地看出对各种表情识别结果的分布情况，对比结果如表 2－表 5 所列。从表 2－表 5 可以看出，误判容易出现在愤怒、高兴、惊奇等面部表情之间，这是因为产生这些表情的面部动作幅度都比较大，可能产生相似的面部特征。在愤怒和悲伤两个表情之间发生误判的可能性也比较大，这是因为这两个表情同属于负面情绪，可能存在相似的面部动作。另外，在几个数据集下恐惧表情的识别准确率都比较低，这是因为恐惧情绪的表情不仅面部动作大，而且不同的人在恐惧时可能有不同的面部反应，从而导致误判发生的可能性较大。

表 2 CK＋数据库中表情识别情况

Table 2 Expression recognition on CK＋ database

（单位：%）

	An	Co	Di	Fe	Ha	Sa	Su
	100	0	0	0	0	0	0
Co	0	95.15	0	0	0	4.85	0
Di	0	0	100	0	0	0	0
Fe	0	0	0	92	2	0	6
Ha	0	0	0	0	100	0	0
Sa	7.76	0	0	0	0	92.24	0
Su	0	0.98	0	0	0	0	99.02

表 3 Oulu-CASIA 数据库中表情识别情况

Table 3 Expression recognition on Oulu-CASIA database

（单位：%）

	An	Di	Fe	Ha	Sa	Su
An	73.6	16.25	1.25	0.15	8.75	0
Di	21.25	76	2.75	0	0	0
Fe	4	1.25	77.83	3.42	2.5	11
Ha	0	0	5.5	90.35	4.15	0
Sa	12.25	0	2.5	0	85.25	0
Su	0	0	7	2	0	91

表 4 MMI 数据库分表情识别情况

Table 4 Expression recognition on MMI database

（单位：%）

	An	Di	Fe	Ha	Sa	Su
An	62.31	25.8	0	0	11.89	0
Di	13.62	74.88	0	8.37	0	3.13
Fe	10.71	0	46.34	10.71	12.29	19.95
Ha	0	0	3.76	96.13	0	0.11
Sa	13.38	3.13	9.27	0	71.1	3.12
Su	2.5	0	18	2.5	0	77

表 5 FERA 数据库分表情识别情况

Table 5 Expression recognition on FERA database

（单位：%）

	An	Fe	Ha	Sa
An	68.41	10.23	0	21.36
Fe	22.56	45.32	32.12	0
Ha	4.66	10.89	84.45	0
Sa	22.78	21.33	0	55.89

4.4.2 与其他算法的对比

将本文提出的模型与其他模型算法进行对比，结果如表 6 所列。从原始文献中提取 HOG 3D，STM-ExpLet，DTAN，DTGN，3DCNN-DAP[14] 和文献[15]的结果，这是几种比较经典的表情识别算法。从表 6 中可以看出，HOG 3D 算法为了提高效率，同时计算多个方向的梯度矢量，虽然缩短了训练时间，但是也影响了训练精度。3DCNN-DAP 算法将脸部分割为多个部分进行特征提取，它不能考虑全局的情况，这在一定程度上影响了训练精度。其余的几种对比算法在训练前都要对图片进行缩放，使用低像素的图像进行训练，使得部分细节特征有所丢失，降低了识别精度。

表 6 算法精确度对比

Table 6 Algorithmic accuracy comparison

（单位：%）

算法	CK＋	Oulu	MMI	FERA
HOG 3D	91.44	70.63	60.89	—
STM-ExpLet	94.14	74.59	75.12	—
DTAN	91.44	74.38	62.45	—
DTGN	92.35	74.17	59.02	—
3DCNN-DAP	92.40	—	63.40	56.10
文献[15]	—	—	—	55.60
本文模型	96.92	82.34	76.53	63.52

从表 6 中可知，本文提出的模型算法在几个数据库下都要优于其他模型，证明这种构想十分有效。在处理图像之前进行强度归一化有效降低了光亮因素对结果的影响，而对比算法大多未采用归一化预处理，这在一定程度上影响了识别的准确性；另外，HOG 3D，3DCNN-DAP 等算法大多只注重精度或效率中的一种，往往导致另外一方面有所欠缺。本文将随机森林算法和头部姿态估计相结合，能够提取出比其他算法更加丰富的脸部信息，在保证准确率的情况下提高了效率。

结束语 本文针对表情识别问题提出的基于增强头部姿态估计的神经网络模型非常有效。通过强度归一化降低外部环境的影响，并使用 RF-HPE 网络使得脸部标志的位置确定更加准确高效。从而可以提取出良好的局部特征。实验结果表明，本文提出的模型在人脸表情识别任务中表现优异，在提高了识别准确率的情况下没有牺牲模型的运行效率，在多个数据集上都取得了不错的效果。同时，本文还存在着一些不足：虽然本模型的效率相比其他算法有所提高，但是算法的实时性效果还不理想，后续拟使用 GPU 并行加速运算，进一步提高算法的实时性，并探究是否能够将其应用在视频任务中。

参 考 文 献

[1] ZOU G F，FU G X，LI H T，et al. A Survey of Multi-pose Face Recognition [J]. Pattern Recognition and Artificial Intelligence，

2015，28(7)：613-625. (in Chinese)
邹国锋，傅桂霞，李海涛，等. 多姿态人脸识别综述[J]. 模式识别与人工智能，2015，28(7)：613-625.

[2] LEVI G，HASSNER T. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns [C]// ACM on International Conference on Multimodal Interaction. ACM，2015：503-510.

[3] WANG D W，ZHOU J，MEI H Y，et al. Summary of facial expression recognition [J]. Computer Engineering and Applications，2014，50(20)：149-157. (in Chinese)
王大伟，周军，梅红岩，等. 人脸表情识别综述[J]. 计算机工程与应用，2014，50(20)：149-157.

[4] ZHANG Z，LUO P，CHEN C L，et al. From Facial Expression Recognition to Interpersonal Relation Prediction [J]. International Journal of Computer Vision，2018，126(5)：550-569.

[5] HAPPY S L，ROUTRAY A. Automatic facial expression recognition using features of salient facial patches[J]. IEEE Transactions on Affective Computing，2015，6(1)：1-12.

[6] GIRSHICK R. Fast R-CNN[J]. Computer Science，arXiv：1504.08083.

[7] ZHOU E，FAN H，CAO Z，et al. Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade [C] // IEEE International Conference on Computer Vision Workshops. IEEE Computer Society，2013：386-391.

[8] KOTSIA I，ZAFEIRIOU S，PITAS I. Texture and shape information fusion for facial expression and facial action unit recognition[C]// First International Conference on Advances in Computer-Human Interaction. IEEE，2008：833-851.

[9] PIETIKINEN M，HADID A，ZHAO G，et al. Computer Vision Using Local Binary Patterns[M]. Springer London，2011.

[10] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[11] LOPES A T，AGUIAR E D，SOUZA A F D，et al. Facial Expression Recognition with Convolutional Neural Networks：Coping with Few Data and the Training Sample Order[J]. Pattern Recognition，2016，61：610-628.

[12] WANI F J，RIZVI S E H，SHARMA M K，et al. A study on cross validation for model selection and estimation[J]. International Journal of Agricultural Sciences，2018，14：165-172.

[13] LIU M，SHAN S，WANG R，et al. Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2014：1749-1756.

[14] JUNG H，LEE S，YIM J，et al. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition[C]// IEEE International Conference on Computer Vision. IEEE Computer Society，2015：2983-2991.

[15] VALSTAR M F，JIANG B，MEHU M，et al. The first facial expression recognition and analysis challenge[C]// IEEE International Conference on Automatic Face & Gesture Recognition and Workshops. IEEE，2011：921-926.