DOI: 10.3979/j.issn.1673-825X.201907030251

多级细节信息融合的人脸表情识别

陈文绪,薛晓军,许江淳,史鹏坤,何晓云
(昆明理工大学 信息工程与自动化学院,昆明 650500)

摘　要: 在自然环境中各种因素的干扰下,人脸表情信息匹配的识别率受到严重影响,针对此问题,提出一种改进的基于 VGGNet16(visual geometry group network16)的网络模型。在 VGGNet16 模型的侧方添加一系列的侧输出层,并在该侧输出层添加不同的卷积核,通过上采样和下采样方法连接侧输出层的上下 2 层,并通过训练使侧输出层能够对其上下 2 层的表情信息进行加权融合。在 VGGNet16 第 5 层的后方添加 2 种不同的卷积核。将侧输出层最终得到的特征图进行局部卷积操作,将 VGGNet16 输出的最终特征图进行全局特征卷积操作,使局部特征与全局特征融合得到最终要进行分类的特征。该模型在 CK+(the extended cohn-kanade)数据集上的识别率为 98.6%,在 RAF-DB(real-world affective faces)数据集上的表情识别率为 79.59%,通过对比常用模型在这 2 种数据集上的识别率发现该模型具有一定的优势。

关键词: 人脸表情识别; 静态图片; 神经网络; 特征融合
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1673-825X(2021)02-0304-07

Facial expression recognition based on multi-level detail information fusion

CHEN Wenxu, XUE Xiaojun, XU Jiangchun, SHI Pengkun, HE Xiaoyun
(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, P. R. China)

Abstract: The recognition rate of facial expression information matching is seriously affected by various factors in the natural environment. Given this problem, the paper proposes an improved network model based on VGGNet16 (visual geometry group network16). Firstly, a series of side output layers are added on the side of the VGGNet16 model, and the different convolution kernels are added in this side output layer, and then the upper and lower layers of the side output layer are connected by oversampling and undersampling, and through training, the side output layer can perform weighted fusion on the expression information of the upper and lower layers. Two different convolution kernels are added behind the fifth layer of VGGNet16, and the final feature map outputted by the side output layer executes local convolution operation, and the final feature map outputted by VGGNet16 executes global feature convolution operation. Finally, the local features are combined with the global features to be classified ultimately. The recognition rate of the model on CK+ (the extended cohn-kanade) data set is 98.6%, and the expression recognition rate on the RAF-DB (real-world affective faces) data set is 79.59%. By comparing the recognition rates of common models on these two data sets, the model has certain advantages.

Keywords: facial expression recognition; static picture; neural networks; feature fusion

收稿日期: 2019-07-03　**修订日期:** 2021-01-20　**通讯作者:** 薛晓军　258467274@qq.com

0 引　言

未来人工智能将会越来越受到重视，服务业也会加入人工智能的元素，比如在医学、心理学以及服务性机器人等领域。由于在真实大自然场景下，摄像头的拍摄会受到光线等因素的影响，故而如何使人脸与过曝的背景分离，并分析出此时的人脸表情信息是一个巨大的考验。

早在 2002 年，文献 [1-2] 首次使用独立分量分析法(independent component analysis，ICA) 和特征脸主分量分析法(principal component analysis，PCA) [3] 在美国军方 FERET(face recognition technology) 数据集上对人脸表情进行实验，其识别率分别接近 90% 和 85%。后来文献 [4] 通过线性分类器并结合隐马尔科夫模型法(hidden markov model，HMM) 对人脸面部特征进行学习分类，之后通过该算法在 CK+数据集上进行测试，人脸表情识别率为 90.9%。可看出早期的人脸表情识别方法在准确率方面存在不足。在 2012 年的 ImageNet 之后出现了许多经典的神经网络模型，文献 [5] 使用了 AlexNet 模型进行人脸表情识别，并对该模型进行参数的优化，在 CK+数据集上的 7 种表情的识别率为 94.4%。为了增加网络的深度和广度，且提高研究对象的识别率，文献 [6] 对 VGGNet 模型改进为 FaceNetExpNet 模型，此网络模型在 CK+数据集上的 8 种表情识别率为 96.8%。文献 [5] 采用 GoogLeNet 模型在 CK+数据集上的识别率为 95%，其提出的基于神经网络的 PP-DN(peak-piloted deep network) 方法在 CK+数据集上的表情识别率为 97.3%。

以上神经网络在对研究对象识别时均有一个特点，首先是对研究对象的特征进行提取，其次对所需要的特征进行学习，最后则是对选择后的特征进行分类，这样便形成了一个一个端到端的识别过程。关于特征识别的神经网络一般由 5 个部分组成: 卷积、激活函数、池化、全连接以及分类。卷积在数学中的作用主要用来对输入函数进行加权累加，在神经网络中主要用来对输入图片的特征进行提取，所提取特征的数量和种类取决于卷积核数量和种类。但是卷积操作之后如果不进行激活操作，那么前面的卷积操作就只是一个简单的线性拟合问题。传统的激活函数常用 sigmoid 函数，但是 sigmoid 激活函数只能保证在 0 附近时函数的斜率才会很大，根据梯度调整后参数才符合要求。但远离 0 的部分，斜率变化会趋近于 0，这样会导致梯度消失，进而影响参数的调整。为了解决这些问题，研究者们使用了 Relu 激活函数。由于人脸表情识别需要多种特征，则需要多种卷积核，这样会导致通道数变得很多，最后会大大增加网络的计算量。这时则可以使用池化操作对特征进行降维，同时池化操作还具有特征拥有平移不变形的性质，尽量减少特征在处理过程中的损失。由于全连接层的参数众多，许多学者目前在寻求一个替代全连接层的方案。

基于以上问题，本文提出在 VGGNet16 网络的侧面加入一些侧输出层(监督模块) ，通过上采样的方法放大特征图像，然后将网络从低级到高级进行特征融合，这样能够确保特征复用，进而减少参数量以及特征在卷积过程中的损失，并且缓解了梯度消失等问题，从而提高表情的识别率。

1 基于 VGG 改进的多级细节信息融合算法

1.1 VGG 网络模型的改进

以 VGGNet16 网络结构为主，上采样和下采样操作为辅的神经网络结构，该结构中使用上采样操作中的去卷积方法，从而实现输入任意大小的图片保证输出符合要求，然后再结合下采样操作使本层特征减小一半。具体操作方法如图 1。该方法在减小特征损失的同时，又大幅度减少了网络的参数量。并且使用下采样方法使侧输出层的特征图缩小一半，然后与下一层的侧输出层的特征图进行融合操作，这样能够使上层的特征图的一些特征能够尽可能保留下来。

图 1 中的 C 表示特征融合操作，卷积核后面的数字表示卷积核的个数，2 个卷积核的堆叠表示进行了 2 次卷积操作。D 表示下采样操作，U 表示上采样操作，GLOBAL 表示全局特征卷积操作，并对 GLOBAL 进行 up×32 操作，LOCAL 表示局部特征卷积操作，SCORE 表示对全局特征和局部特征进行加权融合操作。图中的 3×3 和 5×5 等卷积核是为了丰富特征映射，1×1 的卷积核则是为了降低参数量，提高网络的计算效率。使用不同的卷积核来丰富网络的特征，并通过上采样方法(图中的粗箭头) 把下层侧输出层的特征图放大 2 倍，这样使不同级之间的特征图大小保持一致，方便后面的融合操作。其中 2×2 pooling 为池化中的最大池化操作。

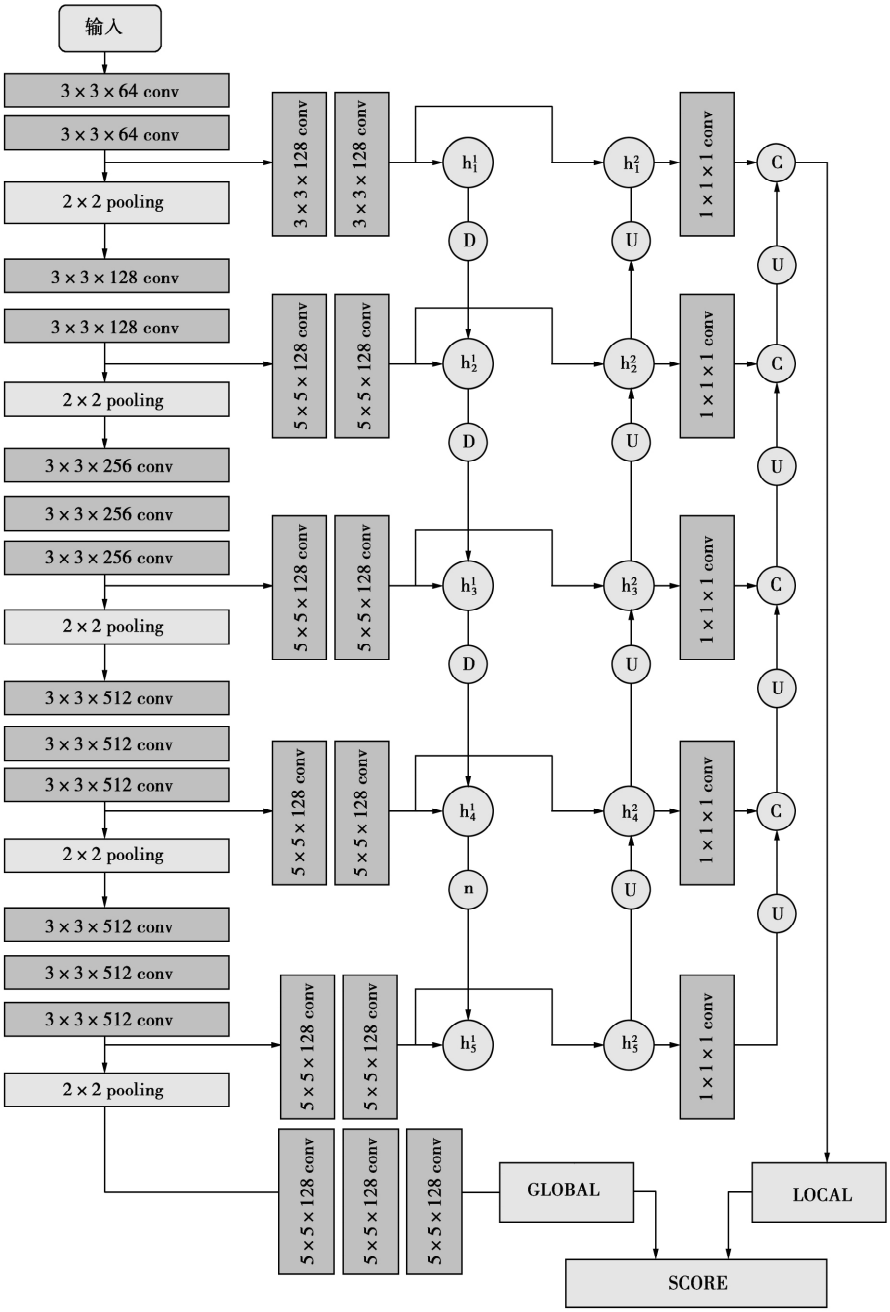


图 1 基于多级细节信息融合方法结构图

Fig.1 Structure diagram based on multi-level detail information fusion method

图 1 第 1 列(VGGNet16) 中,输入的图像经过卷积之后得到的特征图分为 2 路,一路到输出层,另一路进入池化层。进入池化层的这一路经过多次的卷积和池化之后得到的特征图,最终进入 GLOBAL 层进行卷积操作。第 2 列中每个模块有 2 个卷积层,该卷积层的主要作用是进行特征的提取操作,丰富从低级到高级的空间特征信息。第 3、第 4 层的侧输出层都由 2 层卷积核组成,卷积后的特征图一路与上一层下采样操作后的特征图进行融合,另一路则与下一层上采样操作后的特征图进行融合。第 5 层同样通过 2 层卷积核进行卷积。

1.2 融合算法

下采样的融合算法表示为

$$h_i^1 = Down(\varphi(Conv(h_{i-1}^1;\theta_{i-1,i}^1)) + \varphi(Conv(h_i^0;\theta_i^1))) \quad (1)$$

(1) 式中: $Down(\ )$ 表示下采样操作方法; $Conv(\ )$ 表示卷积层; 参数 $\theta=\{W,b\}$; $\varphi(\ )$ 表示该层的激活函数; h_i^1 表示下采样操作后的特征图; h_0^1 的表示第 0

层的特征图参数，初始状态参数为 0。

上采样的融合算法表示为

$$h_i^2 = Up(\varphi(Conv(h_{i+1}^2;\theta_{i,i+1}^2)) + \varphi(Conv(h_i^0;\theta_i^2))) \tag{2}$$

(2) 式中: $Up(\ )$ 表示上采样操作; $Conv(\ )$ 表示卷积层; 参数 $\theta=\{W,b\}$; $\varphi(\ )$ 表示该层的激活函数; h_i^2 表示上采样操作后的特征图值。

之后通过式子 $\varphi(Conv(Cat(h_i^1;h_i^2);\theta_i^3))$ 将前 2 个参数进行融合，将融合后的参数经过 1×1 卷积核的卷积之后再通过上采样方法进行融合，便得到了局部特征。全局特征则是通过使用 2 种不同的卷积核进行卷积。最终使得局部特征与全局特征进行融合。

1.3　参数计算和权重训练

第 l 层侧输出层的激活值用 O_{side} 表示，更新后的侧输出层激活值被定义为 N_{side}。

$$N_{\text{side}} = \begin{cases} \sum_{i=l+1}^{L} w_i^l N_{\text{side}} + O_{\text{side}}, l = 1,\cdots,L-1 \\ O_{\text{side}}, l = L \end{cases} \tag{3}$$

(3) 式中，w_i^1 表示连接侧输出层 i 到侧输出层 l 的权重(i 大于 l) ，当 w_i^1 变为 0 时，则 l 和 i 之间的侧输出层之间的连接则被丢弃。

fusion loss 函数可表示为

$$S_{\text{fuse}}(P,w) = h(\sum_{l=1}^{L} N_{\text{side}}) \tag{4}$$

(4) 式中: $h(\ )$ 表示标准的 softmax loss; P 表示所有标准网络层参数的集合; N_{side} 是第 1 层侧输出的激活值。

测试阶段中融合多个特征映射图，由融合层产生的预测值计算可以表示为

$$P_{\text{fuse}} = \sigma(\sum_{l=1}^{L} N_{\text{side}}) \tag{5}$$

(5) 式中，$\sigma(\ )$ 表示 softmax 函数。

2　实　验

2.1　人脸表情数据集

CK+人脸表情数据集，是一个拥有 2 000 个左右灰度照片的人脸表情数据集。该数据集主要是在实验室中采集的，包含了其中基本的面部表情(中性、蔑视、高兴、悲伤、生气、厌恶、惊讶、害怕) 。

RAF-DB [7] 人脸表情数据集，由北京邮电大学的邓伟洪教授等人于 2017 年建立的。该数据集共收集了 29 672 张从真实的自然界中拍摄的照片，并不仅限于实验室和统一的灰度图像。这些图片中主要有 7 种基本面部表情、复合面部表情和一些研究人脸其他属性的标签图片(人的年龄范围和性别等) ，故而在研究人脸时具有很高的价值。

2.2　软件平台

实验所用电脑 Intel i7，系统 Ubuntu 16.04 Linux，语言 Python，使用的神经网络模型为 Tensorflow 和 Keras，未使用 GPU 加速。

Tensorflow 是 Google 开源的深度学习机构，其具有节省时间，应用广泛，研究团队技术实力强等优点，且该架构支持目前常用的 LSTM(long short-term memory) 、CNN (convolutional neural networks) 和 RNN(recurrent neural network) 等算法。

Keras 神经网络框架比较容易上手，能够堆积 python 中关于 Keras 的库，且能和 Tensorflow 框架配合使用。

2.3　相关参数设置

神经网络训练之前必须对参数进行初始化。为了避免简单的链式计算法在进行一些具有复杂结构的神经网络时不能达到理想的效果，本文将在链式法则中采用随机梯度计算算法。

将权重初始化为接近于 0 的值，即保持初始参数为很小的值。由于当神经元初始参数是随机的但不相等，所以这样可以保证参数在更新后也不相同，近而保证网络的各个部分都不相同。其中权重的更新实现方法表示为

$$W = W - \alpha \frac{\partial}{\partial W} J(W,b) \tag{6}$$

(6) 式中: α 表示训练过程中的学习率; W 表示训练过程中的权重; $J(W,b)$ 表示整体损失代价函数。

对偏置进行初始化，通常会把偏置初始化为 0。其中偏置的更新方法表示为

$$b = b - \alpha \frac{\partial}{\partial b} J(W,b) \tag{7}$$

(7) 式中，b 表示偏置。

对于神经网络中的参数训练，主要参数有: 初始学习速率(initial learning rate，ILR) 、当前迭代次数(iterations) 、批量大小(batch _ size) 、权重衰减(weight decay，WD) 、数据集轮训次数(epoch) 、惯性冲量(momentum) 等。初始时学习速率设置 α 为 0.01，并且在学习过程中通过渐变下降的方法不断更新学习速率的大小。其变化方法表示为

$$v = \beta v - \alpha dx \tag{8}$$

$$x \leftarrow x + v \tag{9}$$

$$\eta(s) = \frac{\eta_0}{1 + s\eta_d} \tag{10}$$

(8)—(10)式中:β 表示中惯性冲量;η_d 表示学习速率衰减;v 表示由 3 个参数量(梯度、惯性冲量、学习速率)决定的幅度;η_0 表示初始学习速率;s 表示当前训练过程中的迭代次数。

对于参数 batch_size 的选定,用改进的网络模型在 CK+数据集进行实验,首先设置最大迭代次数为 50,并保持该值在训练过程中不变,研究 batch_size 的大小对训练总时间和收敛速度的影响,以及保证在指标相同的情况下该神经网络模型找到最优值所需要的迭代次数,并选择 8 个 batch_size 进行实验,实验结果如表 1。由表分析可得,该算法的 batch_size 合适值应为 128 左右。

表 1 关于参数 batch_size 的选定实验

Tab.1 Selected experiment on parameter batch_size

batch_size	训练时间/s	迭代收敛值	最优识别率/%	最优迭代数
4	738	6	97.86	8
8	472	5	97.87	13
16	283	6	97.87	24
32	190	11	98.02	16
64	169	15	98.04	20
128	108	19	98.05	25
256	62	33	98.07	38
512	58	42	98.07	46
1024	50	69	98.07	85

对于 epoch 的选定,在 CK+数据集上进行实验,首先将 CK+数据集的五分之四用来作为训练数据集,五分之一作为测试数据集。然后选定 epoch 为 70,对该数据集进行实验,发现当 epoch 的值为 40 以后训练误差 train loss 几乎没有改变,实验结果如图 2。

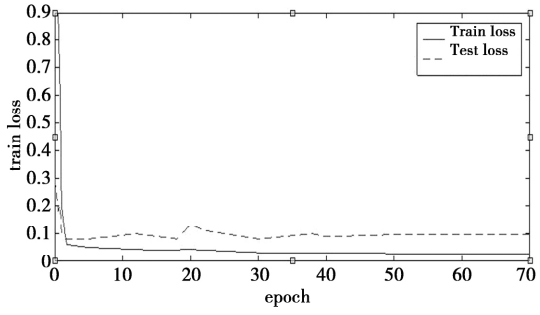


图 2 训练过程中的 Epoch 参数的选取

Fig.2 Selection of Epoch parameters during training

由以上实验结果可得主要参数设置如表 2。

表 2 网络相关参数设置

Tab.2 Network related parameter settings

参数	取值
batch_size	128
WD	0.000 1
epoch	40
ILR	0.01
momentum	0.9

2.4 实验结果

2.4.1 在 CK+人脸表情数据集进行实验

使用本文模型在 CK+人脸表情数据集进行实验,得到验证结果如图 3。图中的数字分别代表 0-生气,1-蔑视,2-恶心,3-害怕,4-高兴,5-悲伤,6-惊讶。图中左侧的 0 到 6 表示样本中含有每种表情的实际样本值,下侧的 0 到 6 表示预测值,预测值表示训练模型在测试样本中测试出的每种表情所占的比例。图 3 中的混合矩阵表示训练模型预测出的每种表情在实际表情样本中所占的比例。通过图 3 中数据可以分析出,生气和蔑视等表情的识别率相对较低。图中颜色越深表示表情识别率越高,颜色越浅表示表情识别率越低。

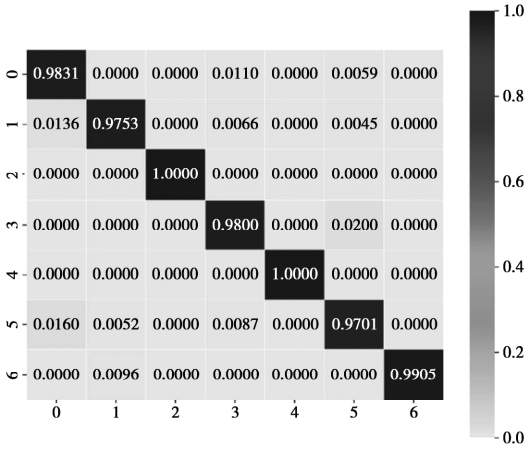


图 3 CK+表情分类混合矩阵

Fig.3 CK+ expression classification hybrid matrix

使用本文改进算法与其他算法在 CK+数据集的实验结果进行对比,结果如表 3。

本文模型与其他模型在 CK+数据集上实验的迭代一次时间对比如表 4。

由表 3 和表 4 可看出本文方法在保证速率的情况下,识别率超过常用的人脸表情识别方法。

表 3　常见模型与本文模型在 CK+数据集上的结果对比

Tab.3　Comparison between common models and the model of this paper on CK+ datasets

实验方法	网络结构	用附加分类器	表情数量	识别率/%
文献 [8]	AlexNet	SVM	7	94.40
文献 [9]	DBM	无	7	91.70
文献 [10]	级联网络	AdaBoost	6	96.70
文献 [11]	级联网络	SVM	8	92.05
文献 [12]	级联网络	SVM	7	93.70
文献 [13]	0 偏置 CNN	无	6	98.30
文献 [14]	FaceNet2ExpNet	无	6	98.60
文献 [15]	Island loss	无	7	94.35
文献 [16]	多任务网络	无	7	95.37
文献 [17]	Clusters loss	无	7	97.10
本文方法	改进 VGG 网络	无	7	98.60

表 4　常见模型与本文模型时间对比

Tab.4　Time comparison between common models and the models in this paper

网络结构	训练时间/s
AlexNet	0.146
GoogLeNet	0.3
本文方法	0.165

2.4.2　在 RAF-DB 数据集进行实验

使用本文模型在 RAF-DB 数据集进行实验，得到实验结果的表情识别率如图 4。

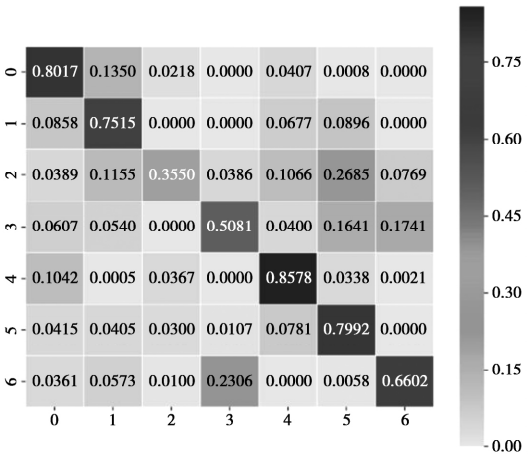


图 4　RAF-DB 表情分类混合矩阵

Fig.4　RAF-DB expression classification hybrid matrix

通过图 4 中数据可以看出，高兴的准确率比较高，恶心和害怕在幅度较浅时比较类似，故这几种表情的识别率较低。

对比本文算法在 RAF-DB 数据集上的准确率与目前主流算法在 RAF-DB 数据集上的识别率，结果如表 5。

表 5　常用算法与本文算法识别率对比

Tab.5　Comparison of commonly used algorithms and algorithm recognition rate

网络结构	表情数量	识别率/%
文献 [7] VGG	7	58.22
文献 [18] MRE-CNN	7	76.73
文献 [7] DLP+SVM	7	74.2
本文方法	7	79.59

通过以上对比分析可以发现，本文的算法在人脸表情识别方面具有一定的优势。

3　结束语

本文提出基于 VGGNet16 改进的多级细节信息融合的人脸表情识别方法，选择 2 种人脸表情数据集 CK+(在实验室中采集的) 和 RAF-DB(来自于真实的大自然中)，通过分析和查阅文献选择合适的网络参数，使用本文方法在这 2 种人脸表情数据集上进行实验，使得出的结果与常用方法在数据集上的实验结果进行比对。最终实验结果表明，本文所提出的方法能够提高人脸表情匹配的识别率。但是，目前的算法对实验的硬件和软件环境有一定的要求，且缺少与微表情相关的数据集，所以下一步应该试着解决这些问题，更加丰富对人类表情的研究。

参考文献:

[1] BARTLETT M S, MOVELLAN J R, SEJNOWSKI T J. Face recognition by independent component analysis [J]. IEEE Transactions on Neural Networks, 2002, 13(6): 1450-1464.

[2] BARTLETT M S, LADES M H, SEJNOWSKI T H. Independent component representations for face recognition [J]. Face Image Analysis by Unsupervised Learning, 1998(612): 39-67.

[3] YAMBOR W S. Analyzing PCA-based face recognition algorithms: eigenvector selection and distance measure [J]. Empirical Evaluation Methods Vision, 2002, 50(3): 39-60.

[4] YEASIN M, BULLOT B, SHARMA R. From facial expression to level of interest: a spatiotemporal approach [C]//Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2004: 922-927.

[5] ZHAO X, LIANG X, LIU L, et al. Peak-piloted deep network for facial expression recognition [C]// European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016: 425-442.

[6] DING H, ZHOU S K, CHELLAPPA R, et al. Facenet2expnet: Regularizing a deep face recognition net for expression recognition [C]//IEEE International Conference on Automatic Face and Gesture Recognition and Workshops. Washington, USA: IEEE, 2017: 118-126.

[7] LI S, DENG W, DU J P. Reliable Crowdsourcing and deep locality-preserving learning for expression recognition in the wild [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii, USA: IEEE, 2017: 2584-2593.

[8] JEON J, PARK JC, JO Y, et al. Real-time emotion recognition for gaming using deep convolution network features [C]//In Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. New York, USA: ACM, 2016: 1-4.

[9] LI J, LAM E Y. Facial expression recognition using deep neural networks [C]//IEEE International Conference on Imaging Systems and Techniques (IST). Macao, China: IEEE, 2015: 1-6.

[10] LIU P, HAN S, MENG Z, et al. Facial expression recognition via a boosted deep belief network [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014: 1805-1812.

[11] LIU M, LI S, SHAN S, et al. Au-aware deep networks for facial expression recognition [C]//IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Shanghai, China: IEEE, 2013: 1-6.

[12] LIU M, LI S, SHAN S, et al. AU-inspired deep networks for facial expression feature learning [J]. Neurocomputing, 2015, 159(1): 126-136.

[13] KHORRAMI P, PAINE T, HUANG T. Do deep neural networks learn facial action units when doing expression recognition [C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. Santiago, Chile: IEEE, 2015: 19-27.

[14] DING H, ZHOU S K, CHELLAPPA R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition [C]//Automatic Face and Gesture Recognition. Washington, USA: IEEE, 2017: 118-126.

[15] CAI J, MENG Z, KHAN A S, et al. Island loss for learning discriminative features in facial expression recognition [C]//IEEE International Conference on Automatic Face and Gesture Recognition. Xi'an, China: IEEE, 2018: 302-309.

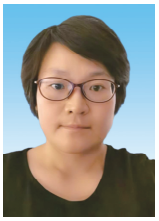
[16] MENG Z, LIU P, CAI J, et al. Identity-aware convolutional neural network for facial expression recognition [C]//Automatic Face and Gesture Recognition. Washington, USA: IEEE, 2017: 558-565.

[17] LIU X, KUMAR B, YOU J, et al. Adaptive deep metric learning for identity-aware facial expression recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Hawaii, USA: IEEE, 2017: 522-531.

[18] JACQUELINE C K L, VITCTOR O K L. Multi-region ensemble convolutional neural network for facial expression recognition [C]//International Conference on Artificial Neural Networks. Rhodes, Greece: Springer, 2018: 84-94.

作者简介:

陈文绪(1994-),男,甘肃定西人,硕士研究生,主要研究方向为表情识别、图像处理。E-mail: 1170365602@qq.com。



薛晓军(1984-),女,山西运城人,实验师,主要研究方向为嵌入式系统开发及应用、图像处理。E-mail: 258467274@qq.com。

许江淳(1962-),男,云南大理人,副教授,主要研究方向为嵌入式系统开发与应用、智能控制。E-mail: jx19631018@163.com。

(编辑: 陈文星)