文章编号：2095-2163(2023)05-0058-07 中图分类号：TP391.4 文献标志码：A

基于深度可分离卷积的表情识别改进方法

李嘉乾，张 雷
（江苏理工学院 电气工程学院，江苏 常州 213001）

摘 要：针对传统表情识别存在相似表情识别精度不高，且深度学习模型参数量巨大问题，提出一种改进的残差网络模型。通过引入深度可分离卷积核，减少了模型的参数量；引入压缩激励模块，改善了模型通道的加权关系；通过将中心损失引入联合算法设计中，提高了相似表情之间的区分度。实验结果表明，识别算法提升了相似表情的区分精度，且较好的控制了模型的参数量。模型在 3 个公开数据集上的准确率分别达到了 97.57%、96.24%、94.09%。

关键词：人脸表情识别；残差网络；深度可分离卷积；压缩激励模块；中心损失

High-precision expression recognition method for complex illumination

LI Jiaqian, ZHANG Lei
(Institute of Mechanical Engineering, Jiangsu University of Technology, Changzhou Jiangsu 213001, China)

【Abstract】 To solve the problems of low accuracy and large number of deep learning model parameters in traditional facial expression recognition, an improved residual network model was proposed. The depth separable convolution kernel was introduced to reduce the number of model parameters, and the compression excitation module was introduced to improve the weighted relationship of model channels. The center loss was introduced into the design joint algorithm to improve the degree of discrimination between similar expressions. The experimental results show that the recognition algorithm improves the discrimination accuracy of similar expressions and controls the number of parameters in the model well. The accuracy of the model on three public data sets is 98.17%, 97.22% and 94.09%.

【Key words】 facial expression recognition; residual network; depthwise separable convolution; squeeze-and-Excitation module; center loss

0 引 言

人工智能在生活中扮演着愈发重要的角色，表情识别是人工智能的一个重要研究方向。Ekman 等[1]把面部表情定义为：厌恶、愤怒、惧怕、愉快、悲伤和惊诧。随着汽车智能化程度的提高，驾驶员面部表情检测已成为比较热门的研究方向[2]。目前，已有对驾驶员进行疲劳驾驶监测与提醒的相关算法[3]。但是，由于传统算法对光照变化的鲁棒性不强，导致光线过亮或光线不充足时，检测不到表情的变化[4]。此外，由于人脸位姿的多变性，使用传统方法检测时，人脸定位需要预先设计人脸提取框[5]，并且由于人脸的照片存在不同的尺度，检测图像时，如果输入人脸的角度发生改变，对最后的精度影响极大[6~8]。

传统人脸表情识别算法是通过手工设计特征提取器进行特征提取，如主成分分析法（Principal Component Analysis，PCA）[9]，局部二值模式（Local Binary Patterns，LBP）[10] 和梯度方向直方图（Histogram of Oriented Gradient，HOG）[11] 等等。然而，传统算法在进行特征提取时，所用的手工特征提取器容易忽略对分类有较大影响的特征信息[12]。而深度学习则不需要人为设计特征提取器[13]，而是通过训练网络结构，用误差反向传播算法不断优化网络参数，使网络自动提取图像特征信息。

Treisman[14] 提出一种模拟人脑注意力机制的模型，其通过计算得到注意力的概率分布结果，从而反应某个输入对于输出的重要作用。目前，在人脸表

基金项目：常州市科技项目（CJ20210070）；江苏省教育厅未来网络科研基金（FNSRFP-2021-YB-35）。

作者简介：李嘉乾（1994-），男，硕士研究生，主要研究方向：微表情识别与网络轻量化；张 雷（1986-），男，博士，副教授，硕士生导师，主要研究方向：认知无线网络与智能无线通信、物联网技术及应用。

通讯作者：张 雷 Email：zhlei@jsut.edu.cn

收稿日期：2022-04-08

情识别领域也受到众多研究者的应用。如：Hu 等[15]提出了基于注意力模块化机制的结合型网络（Squeeze-and-Excitation Networks，SENet）。该网络通过学习的方式，自动获取每个特征通道的重要程度，依照重要程度增强对当前任务重要的特征，并抑制对当前任务用处较小的特征。Li 等[16]提出一种基于注意力机制的自动人脸表情识别网络，该网络将 LBP 特征与注意力机制相结合，增强了注意力模型，获得了更好的效果。

为了提高表情特征的提取能力，同时增强对相似表情的识别能力，提出一种双通道残差网络模型，该模型由两个不同的特征提取网络组成，使之优势互补。对于通道一，本文对 LBP 算子进行改进，在保留其对微小特征敏感性的基础上，进一步提高提取面部纹理特征的能力。但是由于 LBP 方法的定义决定了其关注点更多的是在图像的纹理及轮廓等特征上，在特征提取中侧重方向较为单一，导致提取到表情的微小特征能力强，但相对忽略了与全局的联系。通过增加压缩激励模块，对特征先压缩后进行激发，以提高图像整体的表达能力。将两个通道的特征输入特征融合网络，通过交叉验证方式确定特征融合网络的系数，选择最适合的融合系数以提高网络的分类能力。最后使用 Softmax 函数进行分类，在公开数据集 CK+[17]、Oulu-CASIA[18]和 JAFEE 数据集上进行试验，并与主流算法进行了比较，验证了本文算法的优越性。

1　改进的可分离卷积通道特征网络模型

深度可分离卷积其本质上是将原来的卷积核进行分解，从而实现降低参数量的目的。由于将卷积核拆分，实质上是增加了网络的层数，即增加了网络的深度，有利于网络提取深层特征。以标准的一个深度可分离卷积为例，其总体结构如图 1 所示。

对于卷积层来说，通常情况下一个卷积层内使用的卷积核大小和卷积步长都是相同的，然而深度可分离卷积由于其卷积操作的不同，可以分为两次卷积操作：首先对输入对象进行一次正常的卷积，以此获得每个通道的特征，这也被称为深度卷积；之后通过 1×1 尺寸的卷积核去调整被卷积后的特征通道，并将这些特征融合起来，这也被称为通道卷积。经过两次不同的卷积后，可以大量减少其中的操作量。通常，卷积是全部相乘做全卷积运算，而深度可分离卷积本质上是特征的部分相乘再相加，即深度卷积和通道卷积相加。

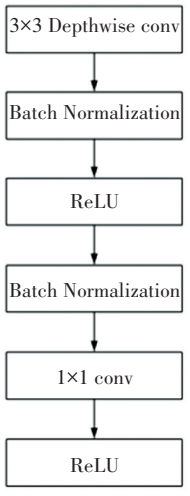


图 1　3×3 大小的深度可分离卷积结构

Fig. 1　3×3 size structure of depth separable convolution

1.1　改进激活函数

神经网络里激活函数的选择是至关重要的，没有激活函数的网络模型，难以处理人脸表情网络输入与输出之间的非线性关系。

通常情况下，激活函数添加在卷积层之后，其作用是增加网络的非线性，以提高网络抗过拟合的能力。目前，使用最多的激活函数是 ReLU，其原理见公式(1)。在正区间，其函数图像是斜率等于 1 的直线，代表输入和输出在正区间都是线性的，并且对函数求导后，其斜率也是不变的，使网络模型保持一个固定的收敛速率，基本杜绝了梯度消失的问题；在负区间，是过原点斜率为 0 的直线，代表此时负区间没有输出。正区间的线性输出和负区间的无输出，组合成了非线性关系。如式(1)

$$\mathrm{ReLU}(\boldsymbol{x}) = \begin{cases} \boldsymbol{x}, & \text{if } \boldsymbol{x} > 0 \\ 0, & \text{if } \boldsymbol{x} \leqslant 0 \end{cases} \tag{1}$$

式中 $\boldsymbol{x}$ 为来自于上一层神经网络的输入向量。

ReLU 激活函数的优点是其结构简单，容易控制收敛速度，但其缺点也显而易见。由于其非线性关系是由正负区间组合而成，对于负区间来说没有输出，与其对应的神经元不在更新参数，相当于这一部分的神经元被舍弃掉了。

本文在 ReLU 激活函数的基础上，提出另一种改进的激活函数，即指数线性单元（exponential linear units，ELR）[19]，其通过对负区间部分进行优化，解决了其负区间神经元不更新参数的问题，并且当输入为负区间时，依然可以保持神经单元的运作性。如公式(2)：

$$\mathrm{ReLU}(\boldsymbol{x}) = \begin{cases} \boldsymbol{x}, & \text{if } x > 0 \\ \delta(e^{x} - 1), & \text{if } x \leqslant 0 \end{cases} \tag{2}$$

其中,参数
δ = 1.673 263 242 354 377 284 817 042 991 671 7。

1.2 引入压缩激发模块

压缩-激发模块(Squeeze-Excitation)本质上属于注意力网络的一种,通过压缩操作和激发操作对通道赋予权重,并依此建立起通道相关的模型,而通道的权重比例依据的是各通道中特征信息的多少来分配的,通过分配权重的多少,判定当前通道与其他通道的优先级关系。而 SE 模块由于其结构中存在池化和激活函数操作,将其放置在每个卷积层之后,可以增大网络的有效感受野,使提取到的特征更能全面的表征图像信息,SE 模块结构如图 2 所示。

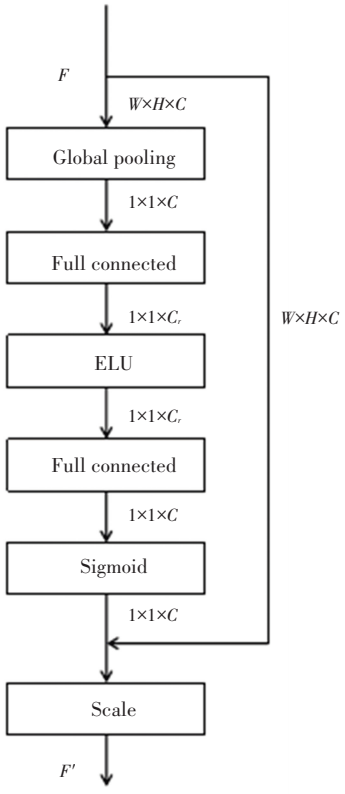


图 2 压缩激励模块结构

Fig. 2 Squeeze-and-Excitation module structure

由上图可以看到,SE 模块主要有 3 个部分组成:分别为 Squeeze (压缩)部分即图中的 Global pooling(全局池化)、Excitation (激发)部分即图中的 sigmoid 激活函数,和 Scale (加权)部分。SE 模块的计算原理是:给其一个输入为特征图,其长宽和维度为 $H \times W \times C$, 经过全局池化后,其维度变成 $1 \times 1 \times C$。 接着,连接两个 FC 层和激活函数层,以增加输出的非线性;之后通过 sigmoid 激活函数,生成一个特征更突出的强特征图。

本文方法的 SE 模块在压缩激发中间使用两个全连接层,其优点在于:

(1)单一的全连接层无法很好的拟合特征通道之间的相关性,对于网络模型非线性的提升起到的作用很小;

(2)由于引入了压缩率,其实是变相降低了网络模型的参数,使得网络可以更快的去判断不同通道之间的重要性。

在 SE 模块的激发部分得到每个特征通道的重要性后,通过输出的强特征经过 Sigmoid 激活函数和原特征加权后,得到该通道的权重值,将其赋予在通道上,就可以实现给通道分配权重。最后,特征通道的增强即是通过加权后得到的每个权重分别乘在对应的通道上来实现。

1.3 交叉熵损失函数

交叉熵损失函数主要刻画的是实际输出与期望输出的距离,也就是交叉熵的值越小,两个概率分布就越接近。假设概率分布 p 为期望输出,概率分布 q 为实际输出,则交叉熵定义如公式(3):

$$H(p,q) = -\sum_{x}(p(x)\log q(x)) \qquad (3)$$

式中:$q(x)$ 表示当前实际的输出概率值,$p(x)$ 表示当前分类值是否是对应对的标签,如果输出值对应标签,则 $p(x)$ 为 1,如果输出值不对应标签,则 $p(x)$ 为 0。其中,$q(x)$ 的值是通过对网络输出的概率分布取对数得到,为的是在不同的标签中更具有区分度,即使得不同样本的样本中心尽可能的互相远离,从而提高表情分类结果的精度。

1.4 改进网络框架

本章节提出了一种结合 SE 模块与可分离卷积的模块以替代网络中的一部分卷积核,并将其修改后嵌入残差网络结构中,如图 3 所示。在图 3 中可以看到一个改进的网络框架,其在本质上是一个轻量化网络,通过将其中一部分卷积核进行分离,从而实现降低模型参数量的目的。表 1 为本文基于深度可分离卷积搭建的网络模型。

表 1 为改进的网络结构及详细参数信息。其中,上表网络中共有 12 层卷积层,其中最开始的两个卷积层使用尺寸为 3×3,步长为 1 的普通卷积;剩余的 10 层为可分离卷积层,其卷积核尺寸有用于深度卷积的 3×3 和 1×1 用于调整通道的,以降低模型的参数量;之后通过最大池化降低特征的 H 和 W 以方便最后的分类;最后使用全局平均池化将输出特征进行相加求和然后取平均值,得出 7 个特征值,将其传入 Softmax 损失函数分类器,对应 7 种表情预测的可能性大小。

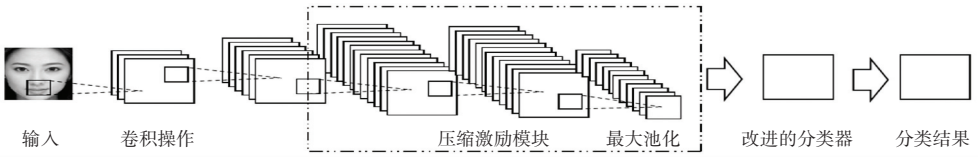


图 3　改进的可分离卷积计算过程

Fig. 3　Improved separable convolution calculation process

表 1　改进的可分离卷积网络结构及参数

Tab. 1　Structure and parameters of improved separable convolution network

网络层数	网络层种类	卷积核大小	卷积步长	输出（$H*W*C$）
1	卷积层	3×3	1	46×46×32
2	压缩激励模块	–	–	46×46×32
3	卷积层	3×3	1	44×44×64
4	压缩激励模块	–	–	44×44×64
5	可分离卷积	3×3	1	44×44×128
6	压缩激励模块	–	–	44×44×128
7	可分离卷积	3×3	1	44×44×128
8	压缩激励模块	–	–	44×44×128
9	最大池化	3×3	2	22×22×128
10	可分离卷积	3×3	1	22×22×256
11	压缩激励模块	–	–	22×22×256
12	可分离卷积	3×3	1	22×22×256
13	压缩激励模块	–	–	22×22×256
14	最大池化	3×3	2	11×11×256
15	可分离卷积	3×3	1	11×11×512
16	压缩激励模块	–	–	11×11×512
17	可分离卷积	3×3	1	11×11×512
18	压缩激励模块	–	–	11×11×512
19	最大池化	3×3	2	6×6×512
20	可分离卷积	3×3	1	6×6×1024
21	压缩激励模块	–	–	6×6×1024
22	可分离卷积	3×3	1	6×6×1024
23	压缩激励模块	–	–	6×6×1024
24	最大池化	3×3	2	3×3×1024
25	可分离卷积	3×3	1	1×1×1024
26	压缩激励模块	–	–	1×1×1024
27	全局池化	–	–	1×1×7

针对全卷积网络模型参数量巨大的问题，本文通过使用可分离卷积替代传统卷积的思路进行优化；本文考虑到虽然可分离卷积可以降低模型参数量，但是过多的堆叠可分离卷积违背了设计的初衷，并且在训练网络的时候发现并不是堆叠可分离卷积就能使模型获得更高的识别精度，过多的可分离卷积反而会使得模型难以训练。所以调节模型结构并设定一个相对合适的网络层数。

2　实验验证与结果分析

2.1　实验环境及数据集介绍

本文所使用环境及计算机配置为 Intel Core i7 8700、32 G 内存、NVIDIA 3060ti 显卡 8 G 显存，软件平台为 Python3.6、TensorFlow－gpu 1.3.1、NVIDIA

CUDA 10.0、cuDNN 7.4.1 库。

为了更好的和其他主流算法比较,本文在对参数调优后,选用 Oulu-CASIA、CK+、JAFEE3 个公共的表情数据集进行实验,各数据集及各表情数量见表 2。

表 2 各数据集表情种类及数量

Tab. 2 Expression types and quantities of each data set

数据集名称	人数	中性	愤怒	厌恶	惧怕	愉快	悲伤	惊诧	总计
Oulu-CASIA	80	8 016	476	462	505	484	453	484	10 880
JAFEE	10	29	30	29	32	31	32	30	213
CK+	70	228	95	79	72	108	110	108	800

3 个数据集及其中样本数量如下:

(1)Oulu-CASIA 表情数据集包含 7 种表情,分别包含厌恶、愤怒、惧怕、愉快、悲伤和惊诧以及中性表情。其中一共包括 10 880 个样本。选取其他 6 种表情样本共 2 864 张,进行数据增广,一共生成了 14 320 张数据集,增广后的数据集样本量为 22 336 张。其中训练集 20 886 张,验证集 1 450 张。

(2)CK+表情数据集同样包含 7 种表情,同样包含厌恶、愤怒、惧怕、愉快、悲伤和惊诧以及中性表情。其中一共包括 800 个样本。进行数据增广,一共生成了 12 000 个样本,其中训练集 10 800 张,验证集 1 200 张;

(3)JAFEE 表情数据集是由日本人和白种人面部情绪图像构成的数据集,包含厌恶、愤怒、惧怕、愉快、悲伤和惊诧以及中性表情。其中一共包括 213 个样本。进行数据增广,一共生成了 10 650 个样本,其中训练集 9 585 张,验证集 1 065 张。

2.2 网络参数设置

本文网络的训练基本参数包含每一批次训练量(Batch-size)、基础学习率(Base-learning rate)、学习率动量(Momentum)、随机失活(Dropout)。网络采用带动量的学习率,将初始学习率设置为 0.01,并采用自适应学习率不断进行修正。考虑到显卡性能及显存,将 Batch-size 设置为 32。Momentum 设为 0.9。为使得模型在训练中减少过拟合现象,并使输出结果具有一定的稀疏性,将 Dropout 设置为 0.5。网络参数设置见表 3。

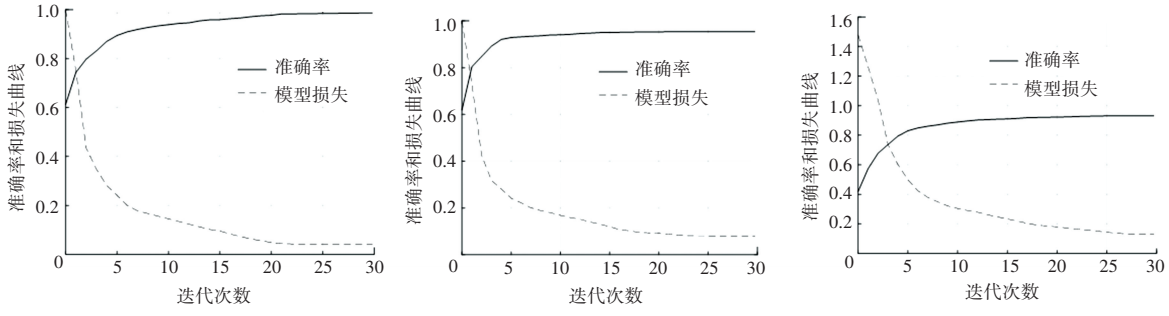
表 3 残差网络参数设置

Tab. 3 Parameters of residual network

参数名称	参数值
一批次训练量	32
基础学习率 μ^t	0.01
总迭代次数	30
初始动量	0.9
随机失活系数	0.5
SE 模块压缩率	4

2.3 实验结果对比分析

在公开数据集 CK+、JAFEE 和 Oulu-CASIA 数据集上用验证集进行实验,经过 30 个 epochs,得到对应的损失(*loss*)和识别率(*accuracy*),分别如图 4(a)、图 4(b)和图 4(c)所示。



(a) CK+数据集实验结果 (b) JAFEE 数据集实验结果 (c) Oulu-CASIA 数据集实验结果

图 4 三个数据集的识别率和损失率

Fig. 4 Recognition rate and loss rate of three data sets

其中,CK+数据集经过 26 个 epochs 后收敛;JAFEE 数据集经过 21 个 epochs 后收敛;Oulu-CASIA 数据集经过 28 个 epochs 后收敛。迭代完 30 个 epochs 后准确率见表 4。

为了验证网络的有效性,本文分别对比了近年来的经典算法和最新算法,其中包括:Alexnet、Inception、Xception、Parallel CNN、CNN、Attention Net、FaceNet2ExpNet、GAN 等,并复现了部分高精度识别网络,且对比了网络之间的模型参数量,对比结果见表 5。

表 4　本文方法识别准确率

Tab. 4　Identification accuracy of this method

数据集	样本数	时间/s	准确率/%
CK+	1 500	19.94	97.57
JAFEE	1 500	19.73	96.24
Oulu-CASIA	1 500	19.91	94.09

表 5　在 CK+数据集上识别率比较

Tab. 5　Comparison of recognition rate on CK+ dataset

采用方法	数据集	时间/s	识别率/%	参数量
Inception[20]	CK+	16.37	94.58	3 807 300
Xception[21]	CK+	114.75	96.30	33 452 455
Parallel CNN[22]	CK+	36.12	96.24	11 000 000
FaceNet2ExpNet[23]	CK+	43.18	97.4	16 691 895
本文方法	CK+	19.94	97.57	5 835 265

由表 5 可见，文献[20]采用了单一的 Inception 结构，其网络层数为 19，参数量较少，但由于其并未对特征提取前端进行预处理，使得特征提取和分类精度完全由网络结构决定，导致了其需要迭代较多次数，才能将网络训练拟合。文献[21]在文献[20]的基础上改进了网络结构，使其分为多个通道进行卷积操作，并将特征图融合，较之前提高了较多的精度，但是由于过多的堆叠了卷积层，使得网络参数巨量增长，模型训练困难，且难以在终端部署。文献[22]在卷积神经网络的主干特征提取网络中作出改进，提高了网络特征提取能力的同时控制了参数量，但由于提取的特征较为单一，导致对于相似表情的识别度不高。文献[23]在 FaceNet 的基础上结合 ExpNet 进行改进，引入滤波对图像进行降噪处理，并根据待检测数据优化网络结构，取得了较高的表情识别精度，但由于其完全使用卷积结构，参数量较大，依赖算力，难以在终端部署。本文引入深度可分离卷积，并在其网络结构上进行优化，使得在保证准确率的情况下，网络参数更少，与文献[21]的基础网络 Xception 相比，由于使用了可分离卷积，网络不需要过多的堆叠卷积层，减少了其卷积层数，使得参数减少了 74%，网络模型的计算复杂度大大降低，符合轻量化网络设计思想。

3　结束语

为了解决传统算法识别精度低且深度学习模型参数量庞大的问题，本文提出了基于深度可分离卷积的残差网络模型。从改进深度可分离卷积中的激活函数入手，提高了模型抗拟合的能力；引入压缩激励模块并设定压缩率，使其提取的特征具有更强的鲁棒性，同时使得提取的结果可以更全面的体现面部表情；在进行表情分类时，通过加入中心损失设计了联合算法，提高了其对类内差异较小的特征的区分能力，即进一步提高了具有相似特征的表情之间的区分度，进而提高了总体表情识别精度。在 3 个数据集（CK+、JAFEE 和 Oulu-CASIA）上分别到达 97.57%，96.24%和 94.09%的识别准确率。实验结果表明，本文提出的改进方案在面部表情识别方面具有很大优势。

参考文献

[1] EKMAN P, FRIESEN W V. Facial action coding system (facs): A technique for the measurement of facial actions[J]. Rivista DI Psichiatria, 2018, 47(2): 126-138.

[2] MA N, WANG Y H. Task complexity analysis method of human-computer interaction in intelligent vehicle cockpit [J / OL]. Journal of graphics, 2022,43(2):356-360.

[3] WANG H T, XIE M D. Research on fatigue driving detection method based on facial features [J]. Journal of Wuhan University of Technology (traffic science and Engineering Edition), 2021, 45 (5): 851-856.

[4] 郭成源. 边缘计算环境下的城市交通道路风险评估与事故风险预测[D]. 南昌：华东交通大学，2021.

[5] CAI Bowen, MA Wufei. Convolutional Neural Network and Bayesian Gaussian Process in Driving Anger Recognition [J]. Engineering, 2020, 12(7):117-126.

[6] 黄忠，胡敏，王晓华. 一种基于几何特征的表情相似性度量方法[J]. 模式识别与人工智能，2015, 28(5):443-451.

[7] 刘鹏. 融合面部表情和语音的驾驶员路怒症识别方法研究[D]. 镇江：江苏大学，2017.

[8] 于申浩. 基于深度学习与信息融合的路怒情绪识别研究[D]. 济南：山东大学，2018.

[9] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 2017, 2(1-3): 37-52.

[10] OJALA T, PIETIK INEN M, HARWOOD D. A comparative study of texture measures with classification based on feature distributions[J]. Pattern Recognition, 1996, 29(1):51-59.

[11] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, IEEE, 2005: 886-893.

[12] EON J W, PARK J C, JO Y J, et al. A real-time facial expression recognizer using deep neural network[C]//Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. New York, NJ, USA: ACM, 2016,94: 1-94.

[13] KIM B Y, ROH J, DONG S Y, et al. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition[J]. Journal on Multimodal User Interfaces, 2016, 10 (2): 173 -189.

[14] Treisman, Anne. How the deployment of attention determines what we see[J]. Vis Cogn, 2006, 14(4-8):411-443.

（下转第 69 页）