工程科学学报
Chinese Journal of Engineering

# 基于S-LRCN的微表情识别算法

李学翰 胡四泉 石志国 张明

# Micro-expression recognition algorithm based on separate long-term recurrent convolutional network

LI Xue-han, HU Si-quan, SHI Zhi-guo, ZHANG Ming

引用本文:
李学翰, 胡四泉, 石志国, 张明. 基于S-LRCN的微表情识别算法[J]. 工程科学学报, 2022, 44(1): 104-113. doi: 10.13374/j.issn2095-9389.2020.06.15.006
LI Xue-han, HU Si-quan, SHI Zhi-guo, ZHANG Ming. Micro-expression recognition algorithm based on separate long-term recurrent convolutional network [J]. *Chinese Journal of Engineering*, 2022, 44(1): 104-113. doi: 10.13374/j.issn2095-9389.2020.06.15.006

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2020.06.15.006

---

工程科学学报, 第 44 卷, 第 1 期: 104−113, 2022 年 1 月
Chinese Journal of Engineering, Vol. 44, No. 1: 104−113, January 2022
https://doi.org/10.13374/j.issn2095-9389.2020.06.15.006; http://cje.ustb.edu.cn

# 基于 S-LRCN 的微表情识别算法

李学翰[1)], 胡四泉[1,2)✉], 石志国[1,2,3)], 张　明[4)]

1) 北京科技大学计算机与通信工程学院, 北京 100083　2) 北京科技大学顺德研究生院, 佛山 528399　3) 北京市大数据中心, 北京 100101
4) 电子科技大学通信与信息工程学院, 成都 611731
✉通信作者, E-mail: husiquan@ustb.edu.cn

**摘　要**　基于面部动态表情序列, 针对静态表情缺少时间信息等问题, 将空间特征与时间特征融合, 利用神经网络在图像分类领域良好的特征, 对需要进行细节分析的表情序列进行处理, 提出基于分离式长期循环卷积网络 (Separate long-term recurrent convolutional networks, S-LRCN) 的微表情识别方法. 首先选取微表情数据集提取面部图像序列, 引入迁移学习的方法, 通过预训练的卷积神经网络模型提取表情帧的空间特征, 降低网络训练中过拟合的危险, 并将视频序列的提取特征输入长短期记忆网络 (Long short-team memory, LSTM) 处理时域特征. 最后建立学习者表情序列小型数据库, 将该方法用于辅助教学评价.

**关键词**　微表情识别; 时空特征; 长期递归卷积网络; 长短期记忆网络; 教学评价
**分类号**　TP391.4

# Micro-expression recognition algorithm based on separate long-term recurrent convolutional network

*LI Xue-han*[1)], *HU Si-quan*[1,2)✉], *SHI Zhi-guo*[1,2,3)], *ZHANG Ming*[4)]

1) School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
2) Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China
3) Beijing Big Data Center, Beijing 100101, China
4) School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
✉ Corresponding author, E-mail: husiquan@ustb.edu.cn

**ABSTRACT**　With the rapid development of machine learning and deep neural network and the popularization of intelligent devices, face recognition technology has rapidly developed. At present, the accuracy of face recognition has exceeded that of the human eyes. Moreover, the software and hardware conditions of large-scale popularization are available, and the application fields are widely distributed. As an important part of face recognition technology, facial expression recognition has been a widely studied subject in the fields of artificial intelligence, security, automation, medical treatment, and driving in recent years. Expression recognition, an active research area in human –computer interaction, involves informatics and psychology and has good research prospect in teaching evaluation. Micro-expression, which has great research significance, is a kind of short-lived facial expression that humans unconsciously make when trying to hide some emotion. Different from the general static facial expression recognition, to realize micro-expression recognition, besides extracting the spatial feature information of facial expression deformation in the image, the temporal-motion information of the continuous image sequence also needs to be considered. In this study, given that static expression features lack temporal information, so that the subtle changes in expression cannot be fully reflected, facial dynamic expression sequences were used

**收稿日期:** 2020−06−15
**基金项目:** 国家自然科学基金资助项目(61977005); 四川省科技计划资助项目(2018GZDZX0034); 北京科技大学顺德研究生院科技创新专项资助项目(BK19CF003); 北京市科技计划资助项目(Z201100004220010)

to fuse spatial features and temporal features, and neural networks were used to provide good features in the field of image classification. Expression sequences were processed, and a micro-expression recognition method based on separate long-term recurrent convolutional network (S-LRCN) was proposed. First, the micro-expression data set was selected to extract the facial image sequence, and the transfer learning method was introduced to extract the spatial features of the expression frame through the pre-trained convolution neural network model, to reduce the risk of overfitting in the network training, and the extracted features of the video sequence were inputted into long short-term memory (LSTM) to process the temporal-domain features. Finally, a small database of learners' expression sequences was established, and the method was used to assist teaching evaluation.

**KEY WORDS** micro-expression recognition; spatial-temporal features; LRCN; LSTM; education evaluation

人脸表情反映了人类的真实情绪，心理学家 Albert Mehrabian 指出“情感表达=7% 语言+38% 声音+55% 面部表情”[1]. 面部表情作为情感和心理的研究载体，在人类情感判断中具有重要的地位. 根据 Ekman 的基本情绪理论，表情包含了大量的情感语义，一般分为高兴、厌恶、愤怒、悲伤、恐惧、和惊讶 6 种[2]. 但是，情感通常是连续的、时序上下文相关的，具有不同的强弱表达关系，基本的情绪理论仍然具有一定的局限性. 与普通表情不同，微表情是在主观情绪影响下产生的一种自发式表情[3]. 微表情具有持续时间短 (1/25 ~ 1/3 s)、动作幅度小等特点[4]，给微表情识别带来了很大的难度.

在以往的微表情识别中通过特征提取的方法对微表情进行分析，但是由于底层特征由人工提取等原因造成特征提取不足，导致微表情识别准确率低[5]. 近年来，深度学习算法表现出强大的优势，尤其是在图像特征提取方面表现突出，准确率远超于传统的特征提取方法[6]. 因此采用深度学习算法来对微表情进行更有效的特征提取以提高识别效果. 此外，传统方法受限于计算能力和表情视频数据的规模，通常使用静态表情或者单表情进行分析，忽略了表情周期性的问题. 表情的产生是一个随时间变化的过程，动态表情更自然地表达了表情变化，而单帧的表情并不能反映表情的整体信息，所以基于动态表情序列进行分析更有助于微表情的识别.

本文基于动态多表情序列，将空间特征和空间时间相结合，提出一种分离式长期循环卷积网络 (Separate long-term recurrent convolutional networks, S-LRCN) 模型，首先将卷积神经网络用于深层特征视觉提取器来提取图像中的微表情静态特征[7]，并将从视频序列中提取的特征提供给由长短期记忆网络 (Long short-team memory, LSTM) 单元组成的双向循环神经网络，得到时序的输出，来提高微表情识别的准确率. 并且研究表情序列的实际使用场景，将教学评价与表情分析结合，通过采集学生面部表情来分析其学习状态，本文采用分心 (Distraction)、专注 (Focus)、疲劳 (Tired) 3 种分类方式建立小型数据库，最后通过改进的 S-LRCN 方法对 3 种状态分类.

## 1 相关工作

### 1.1 表情识别

Ekman 等[8] 于 1976 年提出了面部表情编码系统 (Facial action coding system, FACS). FACS 将人脸区域划分成 44 个运动单元 (Action unit, AU)，并将不同的 AU 进行组合形成 FACS 码，每一种 FACS 码对应着一种面部表情. 并在此基础上，经过对大量表情图片的分析，开发出了面部情感编码系统 (Emotion FACS)[9]. MIT 实验室训练稀疏码本进行微表情的情感分析，通过利用微小时间运动模式的稀疏性，短时间段内在面部和身体区域上提取局部时空特征[10]，从数据中学习微表情码本，并以稀疏方式对特征进行编码，在 AVEC 2012 数据集上的实验表明，这种方式具有很好的性能.

### 1.2 表情特征提取

表情特征的提取方法分为基于静态图像与基于动态图像两类. 其中基于动态特征的提取主要集中在人脸的形变和面部区域的肌肉运动上，基于动态特征提取的代表方法有光流法[11]、运动模型、几何法和特征点跟踪方法等.

Polikovsky 等[12] 通过 3D 直方图的方法，通过关联帧之间的梯度关系进行微表情检测识别. Shreve 等[13] 通过光流法使用应变模式处理长视频，通过在人脸部划分几个特定子区域 (如嘴部，眼睛) 分割面部表情，进而识别微表情. Pfister 等[14] 使用三维正交平面局部二值法 (Local binary patterns from three orthogonal planes, LBP-TOP) 算法提取微表情图像序列的特征，该方法通过二维到三维的扩展提取时域和空域方向上的动态局部纹理特征

进行识别. 梁静等[15]建立 CASME 数据库, 应用 Gabor 滤波提取微表情序列的特征值, 并使用平滑式自适应增强算法结合支持向量机的方法 (Support vector machines based on gentle adaptive boosting, GentleSVM) 建立分类器进行分类识别. Wang 等[16]提出利用 6 交点局部二值方法 (Local binary patterns with six intersection points, LBP-SIP) 对微表情进行识别, 该方法减少了 LBP-TOP 方法中特征的维度, 提高了微表情特征提取的效率.

在基于时空域运动信息描述的微表情识别方面, Liong 等[17]通过利用面部光学应变构造光学应变特征和光学应变加权特征来检测和识别微表情. Le Ngo 等[18]采用欧拉影像放大分析图像频域中的相位以及时域中的幅值, 放大微表情的运动信息, 消除无关的微表情面部动态, 并利用 LBP-TOP 算法进行特征提取. Xu 等[19]提出了一种面部动态映射 (Facial dynamics map, FDM) 的方法来表征微表情序列, 该方法通过计算微表情序列的光流信息然后进行在光流域上的精准对齐.

### 1.3　深度学习与微表情识别

区别于传统的机器学习算法, 深度学习突出了特征学习的重要性, 通过逐层的特征映射, 将原数据空间的特征映射到一个新的特征空间中, 使得分类和预测更加容易. 深度学习可以利用数据提取符合要求的特征, 克服了人工特征不可扩展的缺陷. Patel 等[20]在微表情识别中引入深度学习的方法, 通过特征选择提取微表情特征, 但由于数据集样本量过小, 训练中容易产生过拟合现象, 影响网络的识别准确率. Kim 等[21]使用卷积神经网络对处于不同表情状态的微表情的空间特征进行编码, 将具有表达状态约束的空间特征转移到微表情的时间特征, 使用 LSTM 网络对微表达式不同状态的时间特征进行编码. Khor 等[22]提出一种丰富的长期递归卷积网络, 对数据集提取光流特征以丰富每个时间步或给定时间长度的输入, 该网络通过包括提取空间深层特征和表征时间变化的动态时序模型. Verburg 与 Menkovski[23]通过在微表情图像序列的光流特征上使用递归神经网络, 提取定向光流直方图 (Histogram of oriented optical flow, HOOF) 特征来编码所选面部区域的时间变化, 然后将其传递给由 LSTM 模块以进行检测任务.

## 2　微表情识别方法

微表情识别通过人脸检测算法从复杂场景下获取人脸位置, 检测并分割出人脸轮廓以对其进行微表情的特征提取, 并建立识别分类模型, 其基本步骤包括: (1)人脸表情图像、表情序列的获取与处理; (2)从人脸表情序列中提取微表情特征, 去除特征之间的冗余以降低特征维度; (3)基于长期递归网络, 微表情特征作为时序模型的输入, 用于学习时变输出序列的动态过程; (4)建立动态预测模型, 对人脸微表情分类识别. 如图 1 所示.

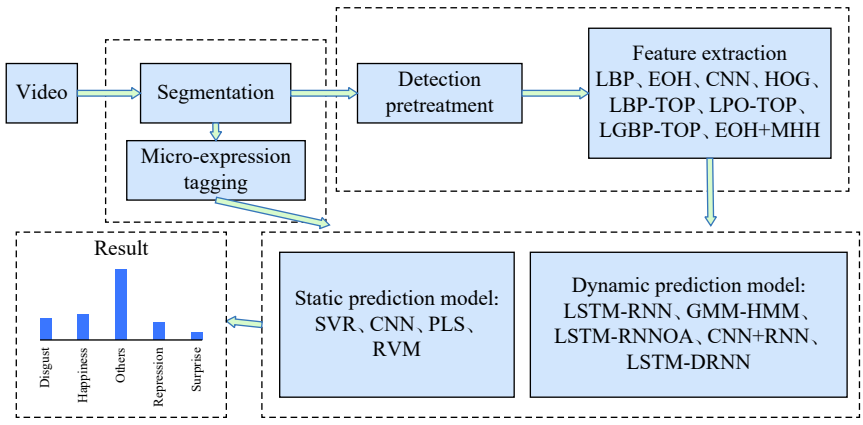


**图 1**　动态表情识别流程

**Fig.1**　Dynamic expression-recognition process

本文方法基于长期循环卷积网络 (Long-term recurrent convolutional networks, LRCN)[7]架构, 并对该模型进行改进使其更适应微表情视频片段的识别, 面对微表情数据集通常存在数据量小的问题, 采用迁移学习的方式避免网络过拟合, 将卷积神经网络 (Convolutional neural networks, CNN) 和 LSTM 的部分微调, 提出 S-LRCN 的方法, 结合卷积神经网络和长期递归网络, 通过两个独立的模块获取空间域特征, 并对时间域特征分类, 首先使用预训练的 CNN 模型提取每一张微表情图片帧

的特征向量组成特征序列，然后将具备时序关联的特征序列输入到 LSTM 网络中，并得到时序的输出. 通过这种方法，可以对 CNN 网络的结构及输出微调，使其分类的准确率更高，并且有利于在小规模数据集上的学习.

### 2.1 LRCN 网络

LRCN 是一种结合传统 CNN 网络和 LSTM 的循环卷积结构[7]，该网络同时具备处理时序视频输入或单帧图片的能力，同时也具备输出单值预测或序列预测的能力，同时适用于大规模的可视学习，LRCN 模型将长期递归网络与卷积神经网络直接连接，以同时进行卷积感知和时间动态学习.

该模型结合深度分层视觉特征提取模型可以学习识别和序列化时空动态任务，包括序列数据（输入、输出）视频，描述等，如图 2 所示. $t$时刻，通过参数化的特征变换将传递给每一个视觉输入 $\boldsymbol{v}_t$（单一图像或视频帧）来产生一个固定长度的矢量$\boldsymbol{l}_t \in \mathbf{R}^d$表示，其中，$\mathbf{R}^d$表示$d$维的实数集，建立视频输入序列的特征空间表示$[\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_3]$，然后输入到序列模型中.

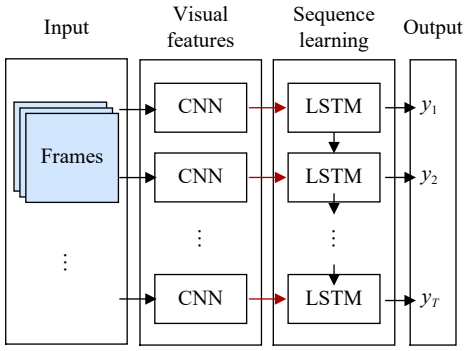


**图 2** LRCN 结构

**Fig.2** LRCN structure

在通常形式下，由序列模型将输入$\boldsymbol{x}_t$和前一个时间步的隐藏状态$h_{t-1}$映射到输出$\boldsymbol{z}_t$和更新后的隐藏状态$h_t$，依次计算$h_1 = f_W(\boldsymbol{x}_1, h_0)$，$h_2 = f_W(\boldsymbol{x}_2, h_1)$，最后得到$h_t$，其中$W$为权值参数. 在时间步$t$预测分布$P(\boldsymbol{y}_t)$的最后一步是在顺序模型的输出$\boldsymbol{z}_t$上取一个softmax逻辑回归函数，将一个向量映射为一个概率分布，产生一个可能的每步时间空间 $C$ 的分布，表示有 $C$ 种结果，$\boldsymbol{y}_t = c$表示第 $c$ 类结果的概率，$W_c$为第 $c$ 类权重向量:

$$P(\boldsymbol{y}_t = c) = \text{softmax}(W_c \boldsymbol{z}_t) = \frac{\mathrm{e}^{W_c \boldsymbol{z}_t}}{\sum_{c \in C} \mathrm{e}^{W_c \boldsymbol{z}_t}} \tag{1}$$

其中，LRCN 针对 3 种主要的视觉问题（行为识别、图像描述和视频描述），实例化的学习任务如下：

1. 顺序输入，固定输出：$[x_1, x_2, \cdots, x_T] \to y$. 面向视觉的行为活动预测，以任意长度$T$的视频作为输入，预测行为对应标签.

2. 固定输入，顺序输出：$x \to [y_1, y_2, \cdots, y_T]$. 面向图像描述问题，以固定图像作为输入，输出任意长度的描述标签.

3. 顺序输入和输出：$[x_1, x_2, \cdots, x_T] \to [y_1, y_2, \cdots, y_T]$. 面向视频描述，输入和输出都是顺序的.

通过实验结果，LRCN 是一种结合空间和时间深度的模型，可以应用于涉及不同维度输入和输出的各种视觉任务，在视频序列分析中具有很好的效果.

### 2.2 S-LRCN 网络

由于微表情是关于视频的帧序列，实现微表情空间域与时间域的特征提取显得尤为重要，所以基于 LRCN“双重深度”序列模型在行为识别中的优势，将 LRCN 用于微表情序列分类，提出一种 S-LRCN 模型. 该方法包含 3 个部分：预处理，微表情特征提取和特征序列分类，其中预处理包括面部裁剪对齐，提取面部关键区域[24]；特征提取包括图片帧预训练面向人脸的 CNN 模型，建立特征集；序列分类将视频序列的特征集提供给由 LSTM 网络，然后分类给定序列是否包含相关的微变化. 该方法具有以下优点：

1. 基于 LRCN，结构简单，需要较少的输入预处理和手工特性设计，减少中间环节；

2. 适合用于微表情数据集数据量不足的情况，通过迁移学习提取面部微观特征，避免训练过程中过拟合；

3. 训练过程可视化，便于修改模型，对参数及特征调优.

S-LRCN 在训练过程中包括两个环节，其中 CNN 用作特征提取器提取表情帧的图像特征，LSTM 用作时序分类器分析特征在时间维度上的关联性.

#### 2.2.1 CNN 作为特征提取器

CNN 作为一种深度学习模型，更适用于提取图像的基础特征并降低模型复杂度，因此采用 CNN 来提取微表情序列的特征向量，在不同环境下的适应性更强，特征表现力更好. 对于微表情识别而言，数据集样本量很小，在网络训练中会出现过拟合的现象，直接从微表情数据训练 CNN 模型是不可行的，为了减少在微表情数据集上训练深度学习网络时的过度拟合，使用基于对象和人脸的 CNN 模型进行迁移学习，使用特征选择来提取与任务相关的深层特征.

Wang 等[25] 在微表情识别中基于迁移学习使

用 ImageNet 数据库初始化残差网络, 并在几种宏观表情数据库上进行进一步的预训练, 最后使用微表情数据集对残差网络和微表情单元进行微调. 但是通常情况下, 宏观表情数据库中的表情变化较大, 具有很明显的表情特征, 而微表情变化幅度小, 更接近没有变化的人脸图像. 因此使用面向人脸识别的 VGGFace 模型[26] 作为微表情帧的特征提取器, 可以从不同环境、人群中提取细微特征, 本文采用的 VGGFace 模型基于通道模型依赖网络 (Squeeze-and-excitation networks, SENet) 架构[27], 并在 VGGFace2 人脸数据库上训练[28]. SENet 通过在残差网络 (Residual network, ResNet)[29] 中嵌入 SENet 结构增强了网络的自适应性, 利用全局信息增强有益特征通道并抑制无用特征通道, 通过特征通道之间的关系提升网络性能. 如图 3 所示.

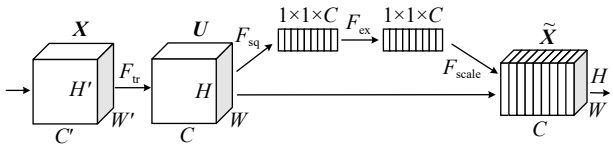


**图 3**　SENet 模块

**Fig.3**　SENet

如图 3, $F_{\mathrm{tr}}: \boldsymbol{X} \to \boldsymbol{U}, \boldsymbol{U}=[u_1,u_2,\cdots,u_k,\cdots,u_C]^{\mathrm{T}}$的实现过程为:

$$u_k=\boldsymbol{v}_k\boldsymbol{X}=\sum_{i=1}^{C'}v_k^i x^i, k=1,...,C \tag{2}$$

其中 $\boldsymbol{v}_k=\left[v_k^1,v_k^2,\cdots,v_k^{C'}\right], \boldsymbol{X}=\left[x^1,x^2,\cdots,x^{C'}\right]^{\mathrm{T}}$, 其中 $v_k^i$ 是二维空间核, $\boldsymbol{v}_k$表示第 $k$ 个卷积核, $x^i$表示第 $i$ 个输入, 经过上述卷积操作后得到特征$\boldsymbol{U}$, 为$W\times H\times C$大小的特征图. 特征压缩将$W\times H\times C$的输入转化为$1\times 1\times C$的输出$z\in\mathbf{R}^C$, 计算如下:

$$z_k=F_{\mathrm{sq}}(u_k)=\frac{1}{H\cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_k(i,j), k=1,...,C \tag{3}$$

特征激发过程得到的特征$\boldsymbol{S}=[s_1,s_2,\cdots,s_C]$的维度是$1\times 1\times C$, 主要用来刻画特征$\boldsymbol{U}$中$C$个特征图的权重, 即:

$$s_k=F_{\mathrm{ex}}(z_k,\boldsymbol{W})=\sigma(g(z_k,\boldsymbol{W}))=\sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1 z_k)), k=1,\cdots,C \tag{4}$$

式中, $\boldsymbol{W}_1\in\mathbf{R}^{\frac{C}{r}\times C}$为全连接层的降维操作, $\boldsymbol{W}_2\in\mathbf{R}^{C\times\frac{C}{r}}$为全连接层的升维操作, 对特征重定向:

$$F_{\mathrm{scale}}(u_k,s_k)=s_k u_k \tag{5}$$

特征提取通过在全局平均池化层 (Global average pooling, GAP) 微调进行特征压缩, 利用两个全连接层去建模通道间的相关性, 并通过减少模型中的参数量和计算量来最小化过度拟合.

### 2.2.2　LSTM 构建序列分类器

由于微表情变化是在连续时间内发生的, 如果没有利用微表情在时间上的信息的话, 很难对微表情变化准确识别. 因此为了利用表情序列在时间上的变化信息, 使用循环神经网络来处理任意时序的输入序列, 可以更容易地处理时间维度信息, 采用 LSTM 节点双向循环神经网络模型处理时序数据, 构建长期递归卷积网络, 对给定序列是否包含相关的微表情判断分类.

定义双向 LSTM 模型的表情特征输入序列 MicroE_Features $=(x_1,\cdots,x_T)$, 前项传播隐变量序列$\overrightarrow{\boldsymbol{h}}=\left(\overrightarrow{h}_1,\cdots,\overrightarrow{h}_T\right)$, 反向传播隐变量序列$\overleftarrow{\boldsymbol{h}}=(\overleftarrow{h}_1,\cdots,\overleftarrow{h}_T)$和输出序列$\boldsymbol{y}=(y_1,\cdots,y_T)$, 则输出序列$\boldsymbol{y}$的更新方式为:

$$\overrightarrow{h}_t=H(\boldsymbol{W}_{\overrightarrow{x}\overrightarrow{h}}x_t+\boldsymbol{W}_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1}+\boldsymbol{b}_{\overrightarrow{h}}) \tag{6}$$

$$\overleftarrow{h}_t=H(\boldsymbol{W}_{x\overleftarrow{h}}x_t+\boldsymbol{W}_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1}+\boldsymbol{b}_{\overleftarrow{h}}) \tag{7}$$

$$y_t=\boldsymbol{W}_{\overrightarrow{h}y}\overrightarrow{h}_t+\boldsymbol{W}_{\overleftarrow{h}y}\overleftarrow{h}_t+\boldsymbol{b}_o \tag{8}$$

式中, $\boldsymbol{W}$为双向 LSTM 模型权重, $\boldsymbol{b}$为偏置项, 偏置项, $H(x)$表示激活函数, 使用长短时记忆神经元进行计算, 双向 LSTM 和记忆神经元如图 4 和 5 所示. 其中图 5 中的$f_t$,$i_t$和$o_t$分别表示遗忘门、输入门和输出门, $C_t$表示记忆单元 (Cell) 在$t$时刻的状态.

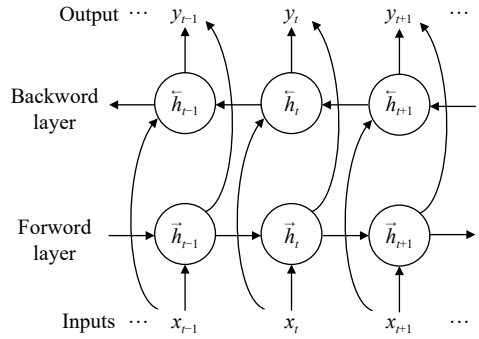


**图 4**　双向循环网络

**Fig.4**　Bidirectional LSTM

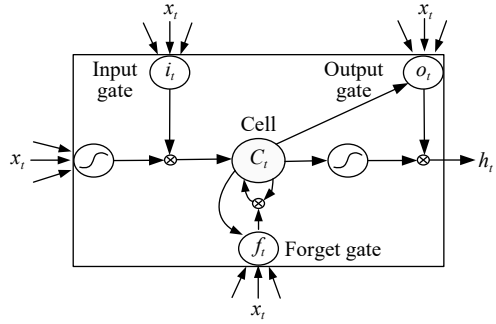


**图 5**　LSTM 神经元

**Fig.5**　LSTM neurons

LSTM 的输入是使用预训练模型从所有序列帧中提取的空间特征，本文采用单层的双向 LSTM 结构，其中包含一个 512 个节点的隐藏层，在 LSTM 隐藏层和全连接层之间使用 Dropout 层以一定概率随机屏蔽神经元，减少神经元间的共适关系，增强网络节点的鲁棒性.

### 2.3 S-LRCN 用于微表情识别

基于以上改进的方法，对于给定的微表情序列，本文实现微表情识别的步骤如下：

（1）载入微表情视频文件，建立序列集 $\boldsymbol{X}=\left(\boldsymbol{X}^1,\boldsymbol{X}^2,\boldsymbol{X}^3,\cdots,\boldsymbol{X}^N\right)$，以及其对应的标签集 $\boldsymbol{Y}=\left(\boldsymbol{Y}^1,\boldsymbol{Y}^2,\boldsymbol{Y}^3,\cdots,\boldsymbol{Y}^N\right)$，$\boldsymbol{X}^i$表示集合中第$i$个微表情序列即$\boldsymbol{X}^i=\left(x_1^i,x_2^i,x_3^i,\cdots,x_{n_i}^i\right)$，$x_j^i$表示第$i$个微表情序列中的第$j$张图片，$n_i$表示第$i$个微表情序列的长度，$\boldsymbol{Y}^i$表示集合中的第$i$个标签.

（2）载入微表情视频文件，首先对序列长度归一化，即输入 LSTM 网络的时间步长设定一个固定值$T$，得到$\boldsymbol{X}^i=\left(x_1^i,x_2^i,x_3^i,\cdots,x_T^i\right)$. 依次对序列归一化的视频序列图片进行人脸检测提取人脸部分，将截取的有效图片尺寸归一化，进而得到处理后的数据集 $\boldsymbol{X}=\left(\boldsymbol{X}^1,\boldsymbol{X}^2,\boldsymbol{X}^3,\cdots,\boldsymbol{X}^N\right)$，此步骤使输入视频序列适合于输入到 CNN 网络.

由于采集的微表情序列含有大量噪声和冗余信息，因此需要去除图像中的无关区域并消除数据噪声，对数据集中的微表情序列进行人脸对齐和人脸剪裁. 使用 Haar 人脸检测器[30]检测人脸，利用主动外观模型 (Active appearance model, AAM) 算法[31]将每个微表情采样序列的中性表情状态下人脸的特征点提取出来，根据特征点坐标裁剪出人脸轮廓，将图像归一化为$224\times224\times3$，避免尺寸差异影响结果.

（3）利用迁移学习和 VGGFace 模型的预训练权重提取面部特征，并对 VGGFace 的预训练权重进行微调，以使模型更有效地适应微表情表达加快收敛，网络输入为大小$224\times224\times3$的人脸表情图像，输出为全局平均池化层之后的全连接层得到的 2048 长度特征向量$\boldsymbol{x}$:

$$\boldsymbol{x}=[m_1,m_2,\cdots,m_n],n=2048 \tag{9}$$

式（9）中，$m_i\in\mathbf{R}^n$，将提取器最后输出的特征向量$\boldsymbol{x}$进行 L2 归一化得到$\bar{\boldsymbol{x}}$:

$$\bar{\boldsymbol{x}}=\frac{m_i}{\sqrt{\sum_{i=1}^n m_i^2}}=\frac{m_i}{\sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}} \tag{10}$$

将最后得到的特征保存到数据集$X=\left\{X^1,X^2,X^3,\cdots,X^N\right\}$，建立特征集，这时$\boldsymbol{X}^i$表示集合中第$i$个提取到的特征序列即$\boldsymbol{X}^i=\left(x_1^i,x_2^i,x_3^i,\cdots,x_T^i\right)$，即针对一个序列生成的$T\times2048$的向量，$x_t^i$表示第$i$个特征序列的第$t$个特征. 将产生的特征向量传递到随后的循环网络中.

（4）由于微表情图像序列具有的动态时域特征，各帧之间包含时域相关性，在完成对微表情单帧图片的空间特征提取之后，利用双向 LSTM 网络前项序列$\overrightarrow{h}$和反向序列$\overleftarrow{h}$传播过程进行训练，获得表情时序特征空间，表情视频序列的每帧人脸图像的表情特征为$x_t\in\mathbf{R}^n$，设定表情变化时序$t\in T$，$T$为表情帧长度，则表情特征时序矩阵为：

$$\boldsymbol{Z}=[\ x_t\quad\cdots\quad x_t\ ] \tag{11}$$

建立顺序输入，固定输出的预测时间分布$[x_1,x_2,\cdots,x_T]\rightarrow y$:

$$y=F(W,\boldsymbol{Z}) \tag{12}$$

式中，$F$为激活函数，$W$为双向 LSTM 的判决参数模型，$y$是多分类的预测结果.

实现步骤如图 6 所示.

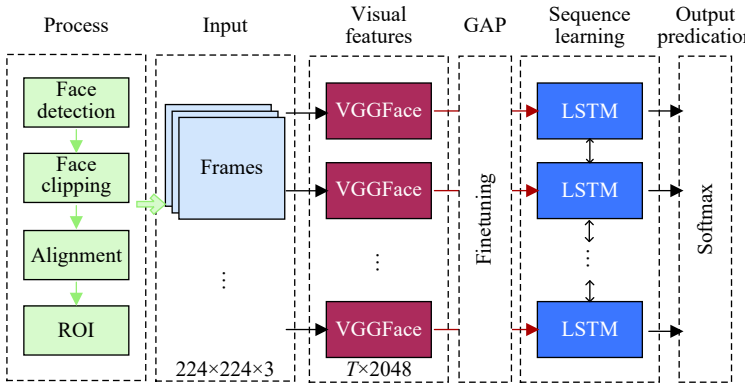


**图 6** 实现方法

**Fig.6** Implementation method

## 3　实验结果

为了验证本文提出的微表情识别方法的性能和准确率, 采用 CASME-Ⅱ数据集进行训练. 首先按照本文的方法训练网络模型, 验证该方法的有效性, 并研究时间序列长度即 LSTM 步长 (Timestep) 以及 LSTM 的深度对模型效果的影响.

### 3.1　数据集选择

采用 CASME-Ⅱ数据集进行实验[32]. CASME-Ⅱ是由中科院心理傅小兰团队所建立的自然诱发的微表情数据库, 包含来自 26 个平均年龄为 22 岁的亚洲参与者的 255 个微表情采样, 视频片段帧数不等. 该数据集在适当的照明条件以及严格的实验环境下采集得到, 图像的分辨率为 640 像素×480 像素. 该数据库样本标有起始帧和结束帧和与之对应的微表情标签, 提供了高兴、厌恶、压抑、惊讶、害怕、伤心及其他情绪分类 (Happiness, surprise, disgust, fear, sadness, repression, others), 数据库中捕捉到的微观表情相对纯粹而清晰, 没有诸如头部动作和不相关的面部动作的噪音. 本文数据集划分为 5 类, 如表 1 所示.

**表 1**　划分情况

**Table 1**　Dataset classification

| Classify | CASME-Ⅱ | Samples |
|---|---|---|
| Happiness | Happiness (32) | 32 |
| Surprise | Surprise (28) | 28 |
| Disgust | Disgust (63) | 63 |
| Repression | Repression (27) | 27 |
| Others | Others (99)<br>Sadness (4)<br>Fear (2) | 105 |

### 3.2　数据集预处理

为了减小不同个体和不同微表情之间的差异, 首先要对数据集中的微表情序列预处理以进行面部对齐, 裁剪得到面部表情区域, 并将图像帧的分辨率统一调整为 224 像素×224 像素, 以便输入空间维度与 VGGFace 网络模型的匹配. 由于数据集中的微表情序列帧数不统一, 针对微表情序列通过时间插值模型插值 (Temporal interpolation model, TIM)[33] 的方法, 将数据集样本每一个图像序列插值为 20 帧, 得到固定长度为 20 的帧序列, 并将 20 帧的序列拆分为两个 10 帧的时间序列, 随后把 10 帧的样本拼接并保存为训练数据, 通过对一段视频的处理获取到两组数据.

由于微表情数据样本数据量较小, 因此对数据集进行扩充, 本文采取镜像模式对数据集进行扩充, 将数据集中的样本逐一进行图片水平镜像, 扩充数据集样本.

### 3.3　实验结果

实验利用 5 折交叉验证的策略, 将数据集随机分为 5 等份, 每一次将其中 4 份作为实验的训练集, 输入到模型中, 另 1 份作为测试集, 用来验证分类的准确率. 网络训练使用早期停止法, 其中将训练集按照 4∶1 的比例随机划分为训练集和验证集. 使用自适应矩估计 (Adam) 优化器, 其中学习率设置为为$10^{-3}$, 衰减为$10^{-5}$, 网络训练为 40 个周期, 批尺寸为 16

选取其中一组训练结果, 当训练趋于稳定时, 自动停止当前训练, 最后得到训练过程中训练集与验证集准确率变化情况, 如图 7 所示.

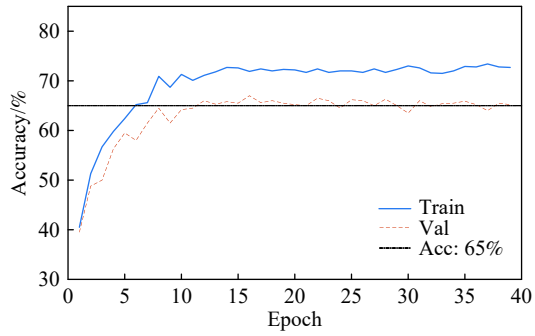


**图 7**　训练曲线

**Fig.7**　Training curve

5 组训练结果如表 2 所示, 得到 5 折交叉验证平均准确率为 65.7%. 最后的分类结果如图 8 所示, 从图中可知, 预测结果在“其他”附近分布比较多, 这是由于 CASME-Ⅱ中将一些无法确定的表情归类到“其他”, 并且此部分数据量相比其他类别较大, 同时实验中将“悲伤”和“害怕”划分到该类表情中, 所以错误的预测结果大多集中在“其他”部分. 如果不考虑“其他”类, 对其他 4 类表情分类会具有更高的准确率.

**表 2**　训练结果

**Table 2**　Training results　　%

| Test1 | Test2 | Test3 | Test4 | Test5 |
|---|---|---|---|---|
| 64.9 | 66.2 | 65.2 | 65.8 | 66.4 |

### 3.4　数据分析

几种微表情识别算法 LBP-TOP[34]、时空完全局部量化模型 (Spatiotemporal completed local quantization patterns, STCLQP)[35]、CNN+LSTM[21]、HOOF+

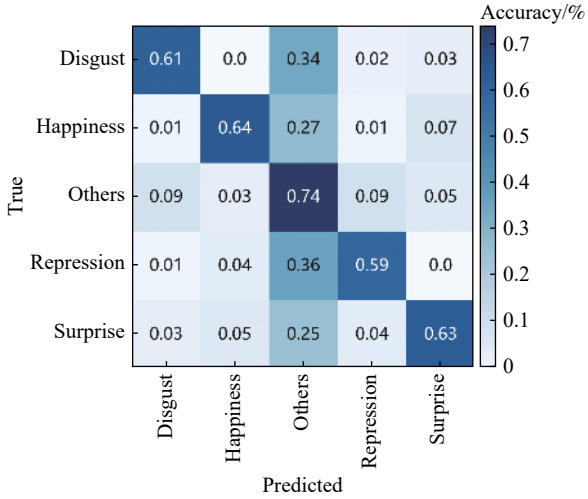


**图 8** 5 种表情分类结果

**Fig.8** Classification results of five expressions

LSTM[23] 及本文研究的 S-LRCN, 采用五折交叉验证的识别准确率对比如表 3 所示, 其中微表情识别算法的数据集采用本文在 CASME-Ⅱ下的分类方法. 通过对比可知, 本文改进的算法对比以往算法识别精度更高, 表示本文算法的可行性. 与传统的机器视觉算法 LBP-TOP、STCLQP 相比, 本文采用深度学习模型在准确率方面提高明显, 并且引入 LSTM 神经元考虑表情变化在时序上的关联特性具有更高的精度; 与 CNN、HOOF 结合 LSTM 的算法相比, 本文通过预训练的卷积神经网络模型提取特征, 采用迁移学习避免网络训练中过拟合的问题, 准确率也有了一定的提高.

**表 3** 不同算法识别准确率

**Table 3** Recognition accuracy of different algorithms

| Methods | Accuracy/% | F1-Score/% |
|---|---|---|
| LBP-TOP | 52.6 | 42.6 |
| STCLQP | 58.6 | 58.0 |
| CNN+LSTM | 61.0 | 58.5 |
| HOOF+LSTM | 59.8 | 56.0 |
| S-LRCN | 65.7 | 60.8 |

基于本文改进的算法, 分别从序列长度、不同 LSTM 模型两个方面来判断这些参数对于 LSTM 模型识别率的影响:

(1)不同长度的微表情序列对识别率的影响, 针对数据集分布采用长度为 6, 10, 15, 30 的 TIM 插值算法, 选择将不同序列的数据输入到单层的双向 LSTM 网络, 实验结果如表 4 所示.

由表 4 可知, 当序列长度较小时, 训练的模型具有更高的准确率, 序列长度为 10 时, 准确率最高为 65.7%, 序列长度为 6 和 15 时, 准确率分别为 62% 和 63.1%. 序列长度为 30 帧时准确率降低到 56.5%, 这是由于微表情通常持续时间很短, 使用短序列可以更快捕捉面部表情的变化情况.

**表 4** 不同序列长度实验效果

**Table 4** Experimental results of different sequence lengths

| Sequence length | Accuracy/% | F1-Score/% |
|---|---|---|
| 6 | 62.0 | 56.6 |
| 10 | 65.7 | 60.8 |
| 15 | 63.1 | 58.6 |
| 30 | 56.5 | 49.6 |

(2)固定序列长度为 10, 分别建立双向 LSTM (512 节点的隐藏层), 2 层双向 LSTM 模型(2 个 512 节点的隐藏层), 单层 LSTM, 多层感知器 (Multi-layer perceptron, MLP), 研究不同 LSTM 模型对识别率的影响如图 9 所示.

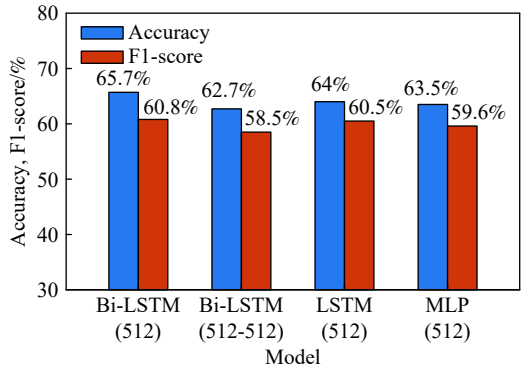


**图 9** 不同 LSTM 模型实验结果

**Fig.9** Experimental results of different LSTM models

由图 9 可知, 使用单层的 LSTM 网络时, 具有更高的精度, 双向 LSTM 为 65.7%, 单向 LSTM 为 64%, 双向 LSTM 识别率更高; 增加隐藏层层数时准确率降低为 62.7%, 这是由于数据量过小, 加深网络深度会导致时间相关性降低; 使用 MLP 网络时训练速度较快, 但是会丢失一些时序特性, 准确率为 63.5%.

实验结果表明, 微表情识别准确率受到序列长度和 LSTM 网络结构的影响, 只有充分考虑网络模型空间特性和时间特性之间的相互关系才能取得更好的效果.

### 3.5 实验扩展

表情分析用途广泛, 将表情识别技术用于教育领域, 通过观察学习者面部表情变化, 分析学习者的心理状态, 从而进一步分析学习者对知识点的理解度及兴趣度等信息, 便于提高教学质量.

基于本文的方法对学习者学习状态进行评价, 采用 CASME-Ⅱ对微表情分类识别, CASME-Ⅱ使用具有情感价值的视频短片来诱发情感表达, 参与者要求在屏幕前观看视频短片, 过程中避免身体运动, 并且在观看短片时保持中立的面部表情试图抑制自己的表情. 由于该数据集在实验室环境下采集, 不易受外界因素干扰, 且视频序列变化微小并不适用于实际的教学场景, 所以建立面向教学评价的小型数据集用于对学习者学习状态的初步评判.

建立模拟教学场景采集人员表情变化, 具体方法如下:

1)选择 30 ~ 45 min 的课程视频片段诱发学习者表情状态, 参与者须观看完整课程视频, 并录制采集视频;

2)参与者观看过程中按一般的上课状态, 头部、肢体动作不做要求;

3)取得的原始数据由参与者去除不相关内容, 筛选表情样本并分类, 表情持续片段为“平静–高峰–平静”的变化区间;

4)筛选的样本由其他参与者对分类结果二次验证, 建立标签.

数据集通过模拟教学场景对参与人员表情变化采集, 参与人员共 6 位, 包含 215 个视频序列, 序列长度为 60 ~ 90 帧, 面部表情标签包括分心、专注和疲惫 (Distraction、focus、tired), 如图 10 所示.


**图 10**　数据分类

**Fig.10**　Data classification

针对建立的教学评价数据集, 采用本文微表中微表情识别方法对学习者学习状态分析, 通过相同的方法建立网络模型, 处理图片序列并划分数据集, 采用五折交叉验证的方法, 验证分类结果的有效性, 取平均值后识别结果如图 11 所示.

## 4　结论

针对目前微表情识别研究中普遍存在的问题展开研究, 通过深度学习来实现对微表情序列的识别分类. 基于 LRCN 在行为识别中优异的性能, 对该方法改进提出一种 S-LRCN 的方法, 该方法更适合用于微表情这种小规模数据集中. 采用迁移学习的方法, 通过预训练的 VGGFace 模型提取表情帧的特征集合以减少数据量过小在训练深度网络中过拟合的风险; 将特征集合输入双向 LSTM 网络以考虑微表情变化持续时间短, 具有时间相关性的特点. 实现表明, 该方法具有较高的准确性. 但是已标记微表情数据量不够, 各类数据分配不均匀以及微表情表现强度普遍较弱仍然是导致识别率低的主要原因, 在以后的研究中还需要进一步完善数据集, 以促进微表情识别的进展.

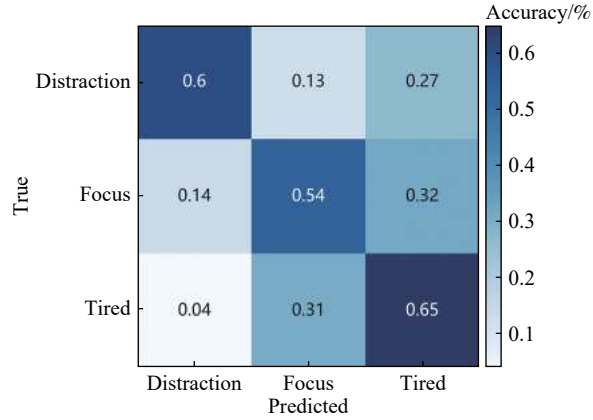


**图 11**　实验结果

**Fig.11**　Experimental result

此外, 将表情识别用于学习场景是构建新型课堂的一种趋势, 基于信息学、心理学和教育学的相关研究基础, 可以通过表情分析研究学习者学习状态. 本文建立了一个包含 3 个类别的小型数据库, 来对教学场景下的表情分类. 今后的工作还要进一步丰富数据, 基于动态表情序列分析学习者情感, 建立心理特征模型, 研究学习过程中学习状态与情感变化的对应关系.

## 参　考　文　献

[1] Mehrabian A. *Nonverbal Communication*. New York: Routledge, 2017

[2] Ekman P. Facial expression and emotion. *Am Psychol*, 1993, 48(4): 384

[3] Ekman P, Friesen W V. Nonverbal leakage and clues to deception. *Psychiatry*, 1969, 32(1): 88

[4] Yan W J, Wang S J, Liu Y J, et al. For micro-expression recognition: Database and suggestions. *Neurocomputing*, 2014, 136(136): 82

[5] Wang S Y. *CNN-RNN Based Micro-Expression Recognition* [Dissertation]. Harbin: Harbin Engineering University, 2018
(王思宇. 基于CNN-RNN的微表情识别[学位论文]. 哈尔滨: 哈尔滨工程大学, 2018)

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 2017, 60(6): 84

[7] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description // 2015 *IEEE Conference on Computer Vision and Pattern Recognition*. Boston, 2015: 2625

[8] Ekman P, Rosenberg E L. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System* (*FACS*). 2nd Ed. USA: Oxford University Press, 2005

[9] Gunes H, Pantic M. Automatic, dimensional and continuous emotion recognition. *Int J Synthetic Emotions*, 2010, 1(1): 68

[10] Song Y L, Morency L P, Davis R. Learning a sparse codebook of facial and body microexpressions for emotion recognition // *Proceedings of the* 15*th ACM International Conference on Multimodal Interaction*. New York, 2013: 237

[11] Li D, Xie L, Lu T, et al. Capture of microexpressions based on the entropy of oriented optical flow. *Chin J Eng*, 2017(11): 1727
(李丹, 解仑, 卢婷, 等. 基于光流方向信息熵统计的微表情捕捉. 工程科学学报, 2017(11): 1727)

[12] Polikovsky S, Kameda Y, Ohta Y. Facial micro-expression detection in Hi-speed video based on facial action coding system (FACS). *IEICE Trans Inform Syst*, 2013, E96-D(1): 81

[13] Shreve M, Godavarthy S, Goldgof D, et al. Macro- and micro-expression spotting in long videos using spatio-temporal strain // 2011 *IEEE International Conference on Automatic Face & Gesture Recognition* (*FG*). Santa Barbara, 2011: 51

[14] Pfister T, Li X B, Zhao G Y, et al. Recognising spontaneous facial micro-expressions // 2011 *International Conference on Computer Vision*. Barcelona, 2011: 1449

[15] Liang J, Yan W J, Wu Q, et al. Recent advances and future trends in micro-expression research. *Bull Natl Nat Sci Foundation China*, 2013(2): 75
(梁静, 颜文靖, 吴奇, 等. 微表情研究的进展与展望. 中国科学基金, 2013(2): 75)

[16] Wang Y D, See J, Phan R C W, et al. LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition // *Asian Conference on Computer Vision – ACCV*2014. Switzerland, 2015: 525

[17] Liong S T, See J, Phan R C W, et al. Spontaneous subtle expression detection and recognition based on facial strain. *Signal Process Image Commun*, 2016, 47: 170

[18] Le Ngo A C, See J, Phan R C W. Sparsity in dynamics of spontaneous subtle emotion: analysis & application. *IEEE Trans Affective Comput*, 2017, 8(3): 396

[19] Xu F, Zhang J P, Wang J Z. Microexpression identification and categorization using a facial dynamics map. *IEEE Trans Affective Comput*, 2017, 8(2): 254

[20] Patel D, Hong X P, Zhao G Y. Selective deep features for micro-expression recognition // 2016 23*rd International Conference on Pattern Recognition* (*ICPR*). Cancun, 2016: 2258

[21] Kim D H, Baddar W J, Ro Y M. Micro-expression recognition with expression-state constrained spatio-temporal feature representations // *Proceedings of the* 24*th ACM international conference on Multimedia*. Amsterdam, 2016: 382

[22] Khor H Q, See J, Phan R C W, et al. Enriched long-term recurrent convolutional network for facial micro-expression recognition // 2018 13*th IEEE International Conference on Automatic Face and Gesture Recognition* (*FG* 2018). Xi'an, 2018: 667

[23] Verburg M, Menkovski V. Micro-expression detection in long videos using optical flow and recurrent neural networks // 2019 14*th IEEE International Conference on Automatic Face and Gesture Recognition* (*FG* 2019). Lille, 2019: 1

[24] Itti L, Koch C. Computational modelling of visual attention. *Nat Rev Neurosci*, 2001, 2(3): 194

[25] Wang C Y, Peng M, Bi T, et al. Micro-attention for micro-expression recognition [J/OL]. *arXiv Preprint* (2019-08-27) [2020-04-21]. https://arxiv.org/abs/1811.02360

[26] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition // *Proceedings of the British Machine Vision Conference* (*BMVC*). Swansea, 2015: 45

[27] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell*, 2020, 42(8): 2011

[28] Cao Q, Shen L, Xie W D, et al. VGGFace2: A dataset for recognising faces across pose and age // 2018 13*th IEEE International Conference on Automatic Face and Gesture Recognition* (*FG* 2018). Xi'an, 2018: 67

[29] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition // 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, 2016: 770

[30] Cootes T F, Edwards G J, Taylor C J. Active appearance models. *IEEE Trans Pattern Anal Mach Intell*, 2001, 23(6): 681

[31] Peng M. *Dual Temporal Scale Convolutional Neural Network for Micro-Expression Recognition* [Dissertation]. Chongqing: Southwest University, 2017
(彭敏. 基于双时间尺度卷积神经网络的微表情识别[学位论文]. 重庆: 西南大学, 2017)

[32] Yan W J, Li X B, Wang S J, et al. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE*, 2014, 9(1): e86041

[33] Zhou Z H, Zhao G Y, Pietikinen M. Towards a practical lipreading system // *The* 24*th IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR* 2011). Providence, RI, 2011: 137

[34] Zhao G Y, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans Pattern Anal Mach Intell*, 2007, 29(6): 915

[35] Huang X H, Zhao G Y, Hong X P, et al. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 2016, 175: 564