

7-15-2020

Learning Expression Recognition Based On Image Sequence

Suqin Wang

1. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China;;

Zhang Feng

1. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China;;

Yudou Gao

2. Information center of Yunnan Power Grid Co., Ltd, Kunming 650200, China;

Shi Min

1. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China;;

Follow this and additional works at: <https://dc-china-simulation.researchcommons.org/journal>



Part of the [Artificial Intelligence and Robotics Commons](#), [Computer Engineering Commons](#), [Numerical Analysis and Scientific Computing Commons](#), [Operations Research](#), [Systems Engineering and Industrial Engineering Commons](#), and the [Systems Science Commons](#)

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

Learning Expression Recognition Based On Image Sequence

Abstract

Abstract: By recognizing the facial expressions, the emotional state of learner can be judged and their learning effect can be analyzed. Due to the persistence and timing of facial expressions, the sequence of facial images is adopted as the object of facial expression recognition. *The Long Short Term Memory Network (LSTM) and VGGNet are combined into a VGGNET-LSTM model.* On this basis, facial expression recognition is carried out, which significantly improves the accuracy of recognition. *Based on the transfer learning method, VGGNet is transferred to the learning expression data set after being pre-trained through the basic expression data set CK+ and avoids the defect of insufficient data in the learning expression data set and solves the problem of overfitting the model.*

Keywords

expression recognition, academic emotion, image sequence, transfer learning

Recommended Citation

Wang Suqin, Zhang Feng, Gao Yudou, Shi Min. Learning Expression Recognition Based On Image Sequence[J]. Journal of System Simulation, 2020, 32(7): 1322-1330.

基于图像序列的学习表情识别

王素琴¹, 张峰¹, 高宇豆², 石敏¹

(1. 华北电力大学 控制与计算机工程学院, 北京 102206; 2. 云南电网有限责任公司信息中心, 云南 昆明 650200)

摘要: 对学习者的面部表情进行识别, 能够判断学习者的情绪状态, 分析其学习效果。针对面部表情具有持续性和时序性的特点, 采用表情图像序列作为表情识别对象。通过组合网络的方式, 将长期记忆网络(Long Short Term Memory Network, LSTM)与 VGGNet 组合成 VGGNet-LSTM 模型, 在此基础上进行表情识别, 显著提高了识别准确率。借鉴迁移学习方法, 将 VGGNet 通过基本表情数据集 CK+ 进行预训练后迁移到学习表情数据集下, 避免了学习表情数据集数据量不足的缺陷, 解决了模型过拟合问题。

关键词: 表情识别; 学业情绪; 图像序列; 迁移学习

中图分类号: TP391.4

文献标识码: A

文章编号: 1004-731X (2020) 07-1322-09

DOI: 10.16182/j.issn1004731x.joss.19-VR0470

Learning Expression Recognition Based On Image Sequence

Wang Suqin¹, Zhang Feng¹, Gao Yudou², Shi Min¹

(1. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China;

2. Information center of Yunnan Power Grid Co., Ltd, Kunming 650200, China)

Abstract: By recognizing the facial expressions, the emotional state of learner can be judged and their learning effect can be analyzed. Due to the persistence and timing of facial expressions, the sequence of facial images is adopted as the object of facial expression recognition. The Long Short Term Memory Network (LSTM) and VGGNet are combined into a VGGNET-LSTM model. On this basis, facial expression recognition is carried out, which significantly improves the accuracy of recognition. Based on the transfer learning method, VGGNet is transferred to the learning expression data set after being pre-trained through the basic expression data set CK+ and avoids the defect of insufficient data in the learning expression data set and solves the problem of overfitting the model.

Keywords: expression recognition; academic emotion; image sequence; transfer learning

引言

心理学家 Mehrabian 等^[1]的研究表明, 人类情感的表达和传递主要通过面部表情、语言、声调等方式, 其中面部表情所传递的信息量高达

55%, 是人类情感表达和传递的重要途径。同语言、姿态相比, 面部表情传达的情感更直接、更丰富、更真实, 通过观察和分析人们的面部表情, 可以推测其情绪状态和心理活动。在学习场景下, 根据学习者的面部表情可以分析出学习者当前的情绪, 发现教学内容或教学方式等方面存在的问题, 并做出针对性的调整。

著名心理学家 Ekman 等^[2]认为人类的情感由离散的基本情绪组成, 基本情绪包括高兴(joy)、悲



收稿日期: 2019-08-30 修回日期: 2019-12-12;
基金项目: 国家自然科学基金(61972379);
作者简介: 王素琴(1970-), 女, 辽宁, 硕士, 副教授, 硕导, 研究方向为计算机视觉, 软件工程; 张峰(1996-), 男, 山西, 硕士生, 研究方向为计算机视觉, 机器学习。

<http://www.china-simulation.com>

• 1322 •

伤(sadness)、厌恶(disgust)、生气(anger)、惊奇(surprise)、恐惧(fear) 6 种。基本情绪与生俱来, 始终贯穿着人们的生活。学业情绪是指学生在学习过程(听课、自习、考试等)中的情绪体验, 与学生的学习情境和学业成就相关, 具有情境性, 多样性和动态性的特征。由于学业情绪是特定情境下的情绪, 因此不能直接套用基本情绪的分类。

国内外学者对学业情绪做了大量的研究。Pekrun 等^[3]将学业情绪划分为希望、愉快、自豪、厌烦、失望、放松、气愤、焦虑、羞愧; 苏世将等^[4]在对大学生的学业情绪研究中将学业情绪分为焦虑、兴趣、气愤、失望、厌烦、放松、愉快、自豪、希望和羞愧 10 种。Sun B 等^[5]对线上教育和线下课堂中的学习者进行了深入的研究, 将学业情绪分为高兴、惊讶、困惑、专注、疲劳、厌烦和分心 7 种, 并采集了学业情绪数据集 BNU-LSVED。

本文提出基于图像序列的学习表情识别, 构建了 VGGFace-LSTM 组合模型。通过 VGGFace 模型提取图像序列中每张图像的视觉特征, 使用 LSTM 模型提取图像序列的时序特征, 在此基础上进行表情分类。

1 相关工作

由于深度学习在图像分类与识别方面展现了强大的处理能力, 近年来采用深度学习技术进行表情识别的研究成为热点。最初的工作主要集中在

在基本表情的识别上, 文献[6]使用卷积神经网络进行面部视角识别和面部 AU 单元识别, 面部视角识别在数据集 BP4D+上达到 97.7 的准确率。文献[7]采用卷积神经网络进行表情识别, 在 CK+数据集^[8]上识别准确率达 97.81%。文献[9]采用图像预处理技术提取特定于表情的特征, 再基于这些表情特征训练卷积神经网络, 实现了实时面部表情识别, 在 BU-3DFE^[10]和 CK+数据集上分别取得 72.89%和 96.76%的识别精度。

随着研究的深入, 人们开始关注各种情境下的表情识别问题。很多学者开展了学习者表情识别的研究工作。学习者表情数据集一般在教学场景下采集, 受光照、遮挡等外界因素的影响, 导致表情识别准确率较低。文献[11]通过表情识别发现网络教学中学生的不良情绪, 表情识别准确率为 70.3%。文献[12]提出一种基于朴素贝叶斯分类的表情识别和评分方法, 将其用于学生情绪的正负面的分类及评分, 在 FER2013 数据集^[13]上识别准确率为 63%。

上述研究都是基于某一时刻的静态图像进行人脸表情识别。但实际上, 人脸表情是动态变化的, 有一个由弱渐强达到峰值、再逐渐消退归于平静的过程, 如图 1 所示, 人的情感信息也隐藏于连续的表情变化过程中。因此, 以展现表情变化过程的表情图像序列为研究对象, 能够更加有效地进行面部表情识别。



图 1 惊奇表情的动态变化过程
Fig. 1 Dynamic process of surprise expression

在深层神经网络中, 循环神经网络(Recurrent Neural Network, RNN)更适合处理序列数据。RNN 中神经元的状态跟输入神经元和上一时刻自身的状态都有关, 这使得 RNN 可以存储时序信息。但是随着时间的推移, RNN 会慢慢“遗忘”早期输入

的信息, 在更新参数时也很难回溯到早期, 导致 RNN 只能处理不太长的序列, 这个问题被称为长期依赖问题。LSTM^[14]在 RNN 基础上进行改进, 增强了循环神经网络的“记忆力”, 目前被广泛用于处理序列化数据。Chao 等^[15]使用 LSTM 结构单

元存储图像序列，识别人脸表情变化。杨雨浓等^[16]提出一种基于 LSTM+RNN 的动态人脸表情识别方法，结合单图像信息和序列关联信息进行表情判别，准确性和鲁棒性都有所增加。

2 模型构建

2.1 基于图像的学习表情识别模型

基于图像的表情识别模型是基于图像序列模型的基础。基于图像的表情识别过程分为人脸图像预处理、表情特征提取和表情分类 3 个步骤。VGGNet^[17]是由牛津大学提出的应用于图片分类的卷积神经网络模型，它使用比较小的卷积核(3×3)和池化核(2×2)，通过反复堆叠 3×3 的小卷积核和 2×2 的最大池化层的方式来不断加深网络结构，网络的性能也因此得以提升。VGGNet 对其他数据集具有很好的泛化能力，在提取图像抽象特征时具有明显的优势。VGGNet 网络深度从 11~19 层不等，比较常用的是 VGG16 和 VGG19。VGGFace 是在 VGG16 的网络结构上，通过人脸数据集 LFW (Labeled Faces In the Wild)^[18]训练得到的，适用于人脸识别。本文使用 VGGFace 进行静态图像的学习表情识别，模型结构如图 2 所示。VGGFace 与 VGG16 网络结构相同，拥有 5 段卷积，每段卷积内有 2~3 个卷积层，同时每段尾部都会连接一个最大池化层。输入的图片数据经过 5 段卷积提取图像视觉特征后通过 3 个全连接层与一个 Softmax 层进行分类。

由于学习表情数据集的数据量较小，为避免深度神经网络出现过拟合问题以及让模型学习到更多人脸表情特征，本文提出一种基于迁移学习的学习表情识别方法。迁移学习的基本原理是将在一种任务中学习到的特征或经验迁移到另一种新的任务中，一方面可以解决目标任务中数据不足的问题，另一方面也能在一定程度上提升模型的学习能力。首先将 VGGFace 模型在基本表情数据集 CK+ 上进行预训练，然后将预训练好的 VGGFace 模型作为图像视觉特征提取器迁移到学

习表情数据集上，提取学习者的表情特征，再将表情特征输入到后面的分类网络中进行表情判别，得到人脸表情对应的情绪类别，模型结构如图 3 所示。

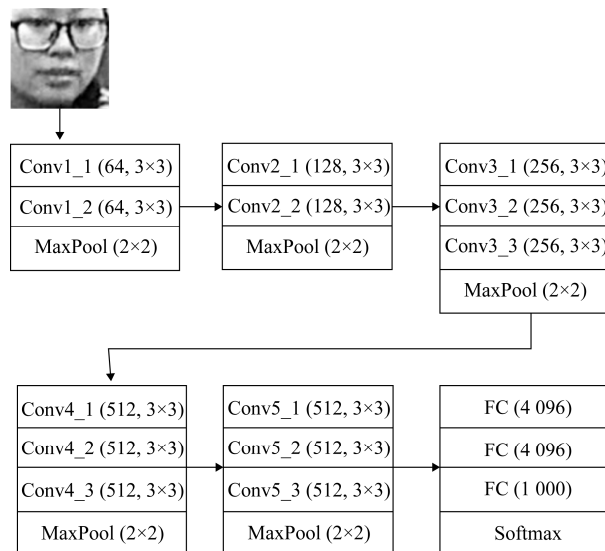


图 2 VGGFace 模型结构
Fig. 2 Model structure of VGGFace

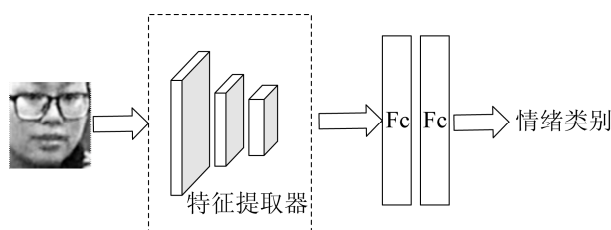


图 3 基于图像的学习表情识别模型结构
Fig. 3 Image-based learning expression recognition model structure

2.2 基于图像序列的学习表情识别模型

构建基于图像序列的学习表情识别模型需要将图像的视觉特征和图像间的时序特征整合起来。本文采用 LSTM 网络提取表情图像序列的时序特征，构建了一个 VGGNet-LSTM 的组合模型，图像序列中的每张表情图片通过 VGGNet 提取视觉特征，然后将 VGGNet 提取的视觉特征按时序顺序输入给 LSTM 中，提取图像序列的时序特征，最后将通过 VGGNet 和 LSTM 提取的特征向量输入到 Softmax 层进行分类，得到图像序列的情绪类别。模型结构如图 4 所示。

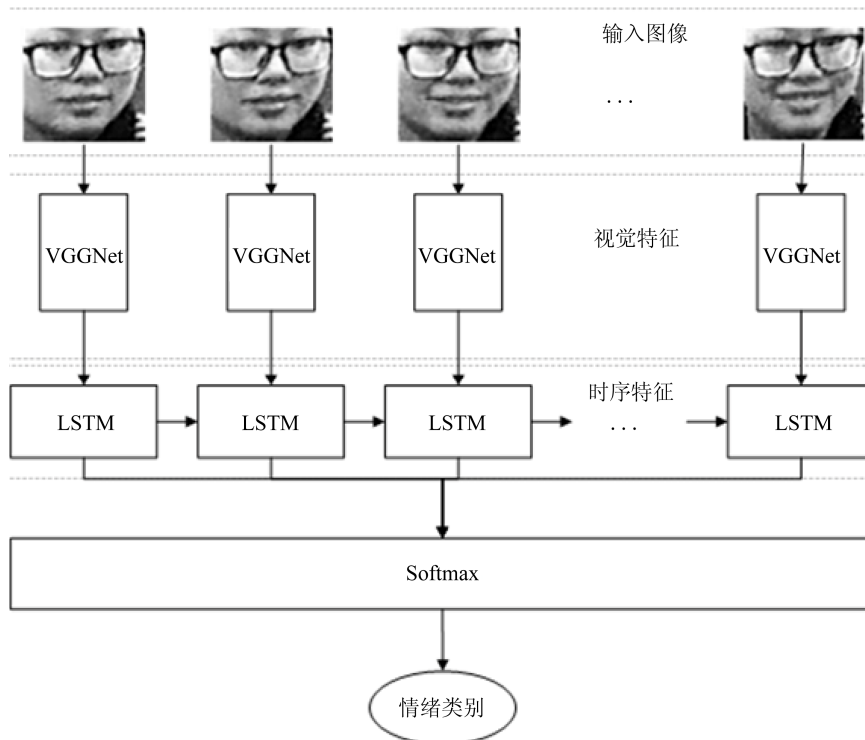


图 4 基于图像序列的学习表情识别模型结构

Fig. 4 Learning expression recognition model structure based on image sequence

将 LSTM 按时间顺序展开, 每个单元具有相同的神经网络结构, 包含了遗忘门、输入门和输出门 3 个门, 每个单元的输出都会传递给下一个, 如图 5 所示。

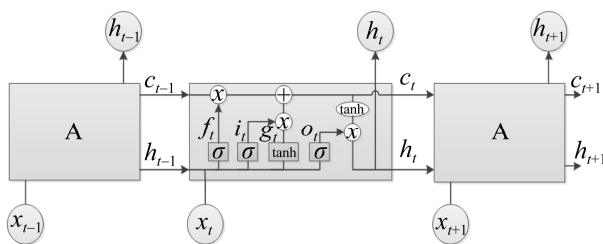


图 5 LSTM 网络结构示意图

Fig. 5 Schematic diagram of LSTM network structure

遗忘门决定从细胞状态中丢弃的信息, 是一个将前一时刻的输 h_{t-1} 出和当前的输入 x_t 作为输入的 sigmoid 函数, 根据式(1)可以计算出上一时刻状态的遗忘比例 f_t 。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

输入门决定当前时刻信息的流入量 i_t , 如式(2)所示。

$$i_t = \text{sigmoid}(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

输出门决定当前时刻的输出比例 o_t , 如式(3)所示。

$$o_t = \text{sigmoid}(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3)$$

g_t 决定当前时刻信息的更新值, 如式(4)所示。

$$g_t = \tanh(W_g \cdot [h_{t-1}, x_t] + b_g) \quad (4)$$

式中: c_t 为当前时刻神经细胞的状态, 根据遗忘门 f_t 和输入门 i_t 的输出对其进行更新; h_t 是输出门 o_t 对当前细胞状态 c_t 进行筛选后的输出, 分别如式(5)~(6)所示。

$$c_t = f_t \times c_{t-1} + i_t \times g_t \quad (5)$$

$$h_t = o_t \times \tanh(c_t) \quad (6)$$

公式(1)~(6)详细介绍了 LSTM 工作机理。

LSTM 神经元之间按时间顺序逐个连接, 每张图像提取的特征都会对下一张图像的特征提取造成影响, 表情图像之间相互作用, 增加了表情识别的可靠性。LSTM 通过输入门、遗忘门和输出门来控制信息流向, 进行信息的筛选, 从而解决了 RNN 的长期依赖的问题。

3 数据集及图像预处理

3.1 数据集

(1) CK+数据集

CK+数据集采集了 123 人共 593 例的动态表情图像序列，每个图像序列包含从表情平静到表情峰值的所有帧，其中仅有 327 例有表情标签。表情图像分为七类，分别为高兴、悲伤、愤怒、厌恶、恐惧、蔑视和惊奇。

本文将每例图像序列中的最后几张接近表情峰值的图像作为样本，经整理得到 3 559 张表情样本。将表情样本中的图像进行镜像、翻转、旋转等处理，最终得到 10 537 张表情图像。随机选择其中的 4/5 作为训练集，其余 1/5 为测试集。

(2) BNU-LSVED 表情数据集

BNU-LSVED 表情数据集采集于真实课堂环境，共有 1 572 例图像序列，每例图像序列约含 40 帧图像。所有图像按学业情绪进行标注，其中专注 196 例，厌恶 205 例，分心 113 例，高兴 402 例，困惑 212 例，惊奇 246 例，疲倦 198 例。疲倦和分心同属于负面的学习情绪，而且两者在面部表情上差异不大，因此将疲倦和分心这 2 种情绪合并，使用分心作为情绪标签。本文基于 BNU-LSVED 构建了 2 个子集，分别为静态图像数据集和图像序列数据集。

在构建静态图像数据集时，将每例图像序列中接近表情峰值的图像作为样本，经预处理后，得到最终的静态图像数据集。数据集共含 2 097 张图像，将其中的 1 687 张划分为训练集，其余作为测试集。

在构建图像序列数据集时，每例图像序列约含 40 张图像，按顺序每 4 张中挑选 1 张，共挑选 10 张图像组成新的图像序列。许多图像序列图像质量差，表情特征不明显，将其剔除。最终的数据集有图像序列 434 例，其中专注 87 例，厌恶 60 例，分心 60 例，高兴 93 例，困惑 82 例，惊奇 52 例。将其中的 356 例作为训练集，其余作为测试集。

3.2 图像预处理

自然场景下拍摄的学习者表情图像会受到光线、角度等因素的影响，因此需要对图像进行预处理。图像灰度化将彩色图像转换为灰度图像，将图像数据量降低为原数据量的三分之一，同时能够降低噪声。灰度化后的图像存在亮度分布不平衡的情况，采用直方图均衡化来增强图像的局部对比度，复现图像中因亮度问题而被忽略的局部细节。图像灰度化与直方图均衡化的处理结果如图 6 所示。

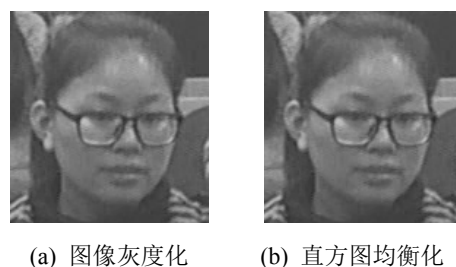


图 6 图像灰度化与直方图均衡化
Fig. 6 Image graying and histogram equalization

人脸图像尺寸不一，为便于后续操作，进行尺度归一化处理。先使用了 OpenFace^[19]的人脸对齐工具对齐人脸，再使用 OpenCV 将人脸图像归一化到 64×64 尺寸，处理结果如图 7 所示。



图 7 尺寸归一化
Fig. 7 Size normalization

4 实验结果分析

本文实验是在 16GB 内存的 NVIDIA GeForce GTX 1080Ti 的 GPU 工作站上进行，通过基于 Tensorflow 后端的 Keras 平台进行模型的构建和训练。

4.1 评判标准

准确率(accuracy)、精确率(Precision)、召回率

(Recall)和 $F1$ 值($F1$ -Measure)是常用的分类评判标准。

准确率是最常见的评价指标,指的是被正确分类的样本数占所有的样本数中的比重。通常来说,正确率越高分类器越好。精度是精确性的度量,表示被分为正例的示例中实际为正例的比例。召回率是覆盖面的度量,度量有多少个正例被分为正例。 $F1$ 值是精确率(precision)和召回率(recall)的调和平均数,最大为 1,最小为 0。计算过程分别如公式(7)~(10)所示。

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$precision = \frac{TP}{TP + FP} \quad (8)$$

$$recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (10)$$

式中: True positives(TP)为被正确地划分为正例的个数; False positives(FP)为被错误地划分为正例的个数; False negatives(FN)为被错误地划分为负例的个数; True negatives(TN)为被正确地划分为负例的个数。

4.2 基于图像的表情识别

4.2.1 基本表情识别

本文选取 VGG16、VGG19、VGGFace 的预训练模型作为视觉特征提取器,在 CK+数据集上进行了表情识别对比实验。VGGFace 模型通过 VGG16 加载 VGGFace 权重进行初始化。模型训练时使用 Nadam 优化器和交叉熵损失函数 categorical_crossentropy 更新和优化模型参数。Nadam 优化器初始学习率设为 0.000 2,模糊因子 epsilon 设为 $1e-08$, beta_1 设为 0.9, beta_2 设为 0.999,学习率衰减值设为 0.004。模型训练时训练集迭代次数为 5 000,在全连接层之后增加 Dropout 层,概率值设为 0.5。VGG16 和 VGG19 模型通过预加载 ImageNet^[20]权重进行初始化,其

余过程与 VGGFace 大致相同。实验结果见表 1。

表 1 CK+数据集上基本表情识别准确率
Tab. 1 Accuracy of basic expression recognition on CK+

data set		%
VGGNet 模型	预训练数据集	识别准确率
VGG19	ImageNet	96.00
VGGFace	LFW	97.57
VGG16	ImageNet	95.24

从实验结果可知,采用 VGGNet 模型进行表情识别可以取得较高的准确率,VGGFace 在人脸数据集 LFW 上进行预训练,与 VGG16 和 VGG19 相比,准确率更高。实验表明 VGGFace 更适合作为表情特征提取器。

4.2.2 学习表情识别

在 Keras 平台下分别加载经过 CK+预训练的 VGGFace 和 VGG19 模型,在 BNU-LSVED 静态图像数据集上进行学习表情识别实验。VGGFace 模型训练时初始迭代次数设为 2 000,但当数据集迭代 200 次时模型已经收敛,后面随着迭代次数的增加准确率并未有所提高反而有下降趋势,因此最终将训练迭代次数设为 200,VGG19 训练过程与 VGGFace 大致相同。实验结果如表 2 所示。

表 2 基于图像的学习表情识别准确率
Tab. 2 Accuracy of face recognition based on image

VGGNet 模型	VGGFace	VGG19
准确率(Accuracy)	56.10	52.44
精确率(Precision)	55.00	58.00
召回率(Recall)	56.00	52.00
F1 值(F1-Measure)	54.00	52.00

由实验结果可得,VGGFace 模型的识别准确率高 VGG19 模型,但 2 种方法在 BNU-LSVED 静态图像数据集上的识别准确率都偏低。使用 VGGFace 和 VGG19 模型进行学习表情识别得到的混淆矩阵如表 3 和表 4 所示。表中第 1 行是表情预测结果,第 1 列是真实情绪标签,表中数值表示模型将人脸图像分类到各个情绪类别的比

例。根据各表情类别分类结果的占比，可以看出每种情绪的分类情况以及混淆情况，表中使用了阴影对每种情绪预测结果占比最高的数值进行了标注。

表 3 VGGFace 学习表情识别的混淆矩阵

Tab. 3 Confusion matrix of expression recognition in VGGFace %

表情	专注	厌恶	分心	高兴	困惑	惊奇
专注	25.5	9.5	36.5	0.0	22.2	6.3
厌恶	5.4	67.6	4.1	2.7	5.4	14.8
分心	10.4	6.0	79.1	0.0	0.0	4.5
高兴	0.0	12.3	1.4	84.9	1.4	0.0
困惑	30.0	4.3	37.1	0.0	14.3	14.3
惊奇	6.5	9.7	14.4	0.0	6.5	62.9

表 4 VGG19 学习表情识别的混淆矩阵

Tab. 4 Confusion matrix of expression recognition in VGG19 %

表情	专注	厌恶	分心	高兴	困惑	惊奇
专注	19.0	3.2	12.7	0.0	60.3	4.8
厌恶	1.3	44.0	1.3	4.0	4.0	45.3
分心	3.0	7.5	37.3	0.0	46.2	6.0
高兴	0.0	16.4	0.0	68.5	1.4	13.7
困惑	10.0	4.3	8.5	0.0	74.3	2.9
惊奇	4.8	9.7	0.0	0.0	16.1	69.4

从表 3~4 结果可知，这 2 个模型对高兴、惊奇两种情绪的识别召回率都较高，这 2 种情绪同时也属于基本情绪，较其他情绪容易区分。VGGFace 模型对专注和困惑 2 种情绪的分类结果较差，VGG19 模型对专注、厌恶、分心 3 种情绪的分类效果较差。相比较而言，除了困惑和惊奇两种情绪 VGG19 识别率比 VGGFace 高以外，其余四种情绪 VGGFace 的识别率都更好，这说明 VGGFace 更适合用于学习表情的视觉特征提取器。

从实验结果上看，专注、分心、困惑这些在学习场景下截然不同的情绪，仅仅通过单张静态人脸图像，难以准确的加以区分。

4.3 基于图像序列的学习表情识别

在 BNU-LSVED 数据集上进行基于图像序列

的学习表情识别实验。首先分别使用预训练的 VGGFace 和 VGG19 模型提取表情图像视觉特征，再分别使用 LSTM 和 RNN 网络提取视觉特征序列的时序特征，最后进行表情分类。为证明通过提取时序特征进行表情识别网络结构的优越性，本文进行 2 次对比试验。实验 1 将 VGGNet 提取的视觉特征进行特征融合后通过 Softmax 层进行表情分类。实验 2 将图像序列数据集转换为静态表情图像数据集，通过 VGGNet 网络进行表情分类。实验结果如表 5 所示。

表 5 基于图像序列的学习表情识别准确率

Tab. 5 Expression recognition accuracy based on image sequence %

表情识别模型	VGGFace	VGG19
VGGNet-LSTM	65.93	62.51
VGGNet-RNN	63.74	60.44
VGGNet 特征融合	61.54	58.93
VGGNet	56.04	52.20

由实验结果可知，相比通过静态图像进行表情识别，经过特征融合进行表情识别的准确率大大提升，但仍然低于通过 RNN 和 LSTM 提取图像序列时序特征后再进行分类的模型的识别率。使用 LSTM 模型的识别率优于使用 RNN 的模型，同时使用 VGGFace 提取学习表情图像视觉特征的网络能获得更高的表情识别准确率。由此证明通过 VGGFace-LSTM 组合模型进行表情识别能获得更高的识别准确率。

VGGFace-LSTM 模型的混淆矩阵如表 6 所示，表 6 中使用阴影对占比最高的数值进行了标注。由表 6 可得，VGGFace-LSTM 模型的学习表情识别准确率较 4.2.2 节中 VGGFace 模型有了很大的提高，尤其是分心和高兴，这 2 种情绪的召回率达到了 90% 以上。对于 VGGFace 模型，专注、分心和困惑这 3 种情绪难以区分，而对于 VGGFace-LSTM 模型，专注和分心都能以 60% 以上的召回率被正确分类。

表 6 基于图像序列的学习表情识别的混淆矩阵
Tab. 6 Confusion matrix of expression recognition based on image sequence %

表情	专注	厌恶	分心	高兴	困惑	惊奇
专注	61.1	16.7	16.7	0.0	5.6	0.0
厌恶	16.7	66.7	0.0	0.0	0.0	16.7
分心	0.0	7.1	92.9	0.0	0.0	0.0
高兴	10.0	0.0	0.0	90.0	0.0	0.0
困惑	12.5	6.3	5.0	0.0	25.0	6.3
惊奇	0.0	0.0	45.5	0.0	0.0	54.5

5 结论

面部表情是一个动态变化的过程, 仅使用某一时刻的静态人脸图像进行表情识别准确率较低。本文提出基于图像序列的学习表情识别方法, 将提取图像视觉特征的 VGGFace 网络以及提取图像序列时序特征的 LSTM 网络组合起来进行学习表情识别。

实验表明, 与基于静态图像的学习表情识别相比, 使用图像序列能获得更高的表情识别准确率。基于图像的模型, 专注、分心和困惑 3 种情绪难以区分, 而基于图像序列的模型, 专注和分心都能以 60% 以上的召回率被正确分类。虽然基于图像序列的学习表情识别对分心和高兴这 2 种情绪的召回率达到了 90% 以上, 但其他 4 种学习表情的召回率都不太高, 尤其对于困惑表情, 更容易被识别为分心。

另外, 与在基本表情数据集 CK+ 上的表情识别相比, 使用真实课堂下采集的学习表情数据集, 无论是通过静态图像或是通过图像序列进行表情识别的识别率都要低很多, 无法满足实际的需要。

未来可以在表情识别的同时, 增加语音识别、身体姿态识别等进行多模态的情感识别, 提高情感识别的准确率。

参考文献:

- [1] Mehrabian A, Russell J A. An approach to environmental psychology[M]. Cambridge: MIT Press, 1980.
- [2] Ekman P, Friesen W V. Constants across cultures in the

face and emotion[J]. Journal of Personality and Social Psychology (S0022-3514), 1971, 17(2): 124.

- [3] Pekrun R, Linnenbrink-Garcia L. Academic Emotions and Student Engagement[M]. Springer US: Handbook of Research on Student Engagement, 2012.
- [4] 苏世将, 马惠霞. 大学生的一般学业情绪现状及其与大学适应的关系研究[J]. 中国健康心理学杂志, 2009, 17(5): 591-593.
Su Shijiang, Ma Huixia. A study on the general academic mood of college students and its relationship with university adaptation[J]. China Journal of Health Psychology, 2009, 17(5): 591-593.
- [5] Sun B, Wei Q, He J, et al. BNU-LSVED: a multimodal spontaneous expression database in educational environment[C]. Optics and Photonics for Information Processing X. California: SPIE - International Society for Optics and Photonics, 2016, 9970: 16-17.
- [6] He J, Li D, Yang B, et al. Multi View Facial Action Unit Detection Based on CNN and BLSTM-RNN[C]. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Washington DC: IEEE Computer Society, 2017.
- [7] Lopes A T, Aguiar E D, Oliveira-Santos T. A Facial Expression Recognition System Using Convolutional Networks[C]. Graphics, Patterns and Images. Salvador: IEEE, 2015: 273-280.
- [8] Lopes A T, de Aguiar E, De Souza A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order[J]. Pattern Recognition (S0031-3203), 2017, 61(4): 610-628.
- [9] Lucey P, Cohn J F, Kanade T, et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression[C]. Computer Vision and Pattern Recognition Workshops. San Francisco, CA: IEEE, 2010: 94-101.
- [10] Yin L, Wei X, Sun Y, et al. A 3D facial expression database for facial behavior research[C]. International Conference on Automatic Face & Gesture Recognition. Southampton: IEEE, 2006.
- [11] 袁源. 远程教育中不良表情的识别研究[D]. 成都: 四川师范大学, 2014.
Yuan Yuan. Recognition of bad expressions in Distance Education[D]. Chengdu: Sichuan Normal University, 2014.
- [12] 唐康. 人脸检测和表情识别研究及其在课堂教学评价

- 中的应用[D]. 重庆: 重庆师范大学, 2019.
- Tang Kang. Research on face detection and expression recognition and its application in classroom teaching evaluation[D]. Chongqing: Chongqing Normal University, 2019.
- [13] Goodfellow I J, Erhan D, Carrier P L, et al. Challenges in representation learning: A report on three machine learning contests[J]. *Neural Networks (S0893-6080)*, 2015, 64: 59-63.
- [14] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A Search Space Odyssey[J]. *IEEE Transactions on Neural Networks & Learning Systems (S1045-9227)*, 2016, 28(10): 2222-2232.
- [15] Chao L, Tao J, Yang M, et al. Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition[C]. *International Workshop. Brisbane, Australia: ACM*, 2015: 65-72.
- [16] 杨雨浓. 基于深度学习的人脸表情识别方法研究[D]. 西安: 西北大学, 2017.
- Yang Yunong. Research on facial expression recognition based on deep learning[D]. Xi'an: Northwest University, 2017.
- [17] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL]. Eprint arXiv: 1409.1556, 2014, <http://arxiv.org/abs/1409.1556>.
- [18] Huang G B, Mattar M, Berg T, et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments[R]. *University of Massachusetts Amherst Technical Report*, 2008: 7-49.
- [19] Baltru Aitis T, Robinson P, Morency L P. OpenFace: An open source facial behavior analysis toolkit[C]. *IEEE Winter Conference on Applications of Computer Vision. Lake Placid, NY: IEEE*, 2016: 1-10.
- [20] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. *2009 IEEE conference on computer vision and pattern recognition. Miami, FL: IEEE*, 2009: 248-255.