# 基于表情及姿态融合的情绪识别

文虹茜，卿粼波，晋儒龙，王　露
(四川大学电子信息学院，成都 610065)

**摘　要**：情绪识别指在使计算机拥有能够感知和分析人类情绪和意图的能力，从而在娱乐、教育、医疗和公共安全等领域发挥作用. 与直观的面部表情相比，身体姿态在情绪识别方面的作用总是被低估. 针对公共空间个体人脸分辨率较低、表情识别精度不高的问题，提出了融合面部表情和身体姿态的情绪识别方法. 首先，对视频数据进行预处理获得表情通道和姿态通道的输入序列；然后，使用深度学习的方法分别提取表情和姿态的情绪特征；最后，在决策层进行融合和分类. 构建了基于视频的公共空间个体情绪数据集(SCU-FABE)，在此基础上，结合姿态情绪识别数据增强，实现了公共空间个体情绪的有效识别. 实验结果表明，表情和姿态情绪识别取得了 94.698%和 88.024%的平均识别率；融合情绪识别平均识别率为 95.766%，有效融合了面部表情和身体姿态表达的情绪信息，在真实场景视频数据中具有良好的泛化能力和适用性.

**关键词**：深度学习；情绪识别；决策层融合；面部表情；身体姿态
**中图分类号**：TP391.4　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.043002

# Emotion recognition based on fusion of expression and posture

*WEN Hong-Qian*, *QING Lin-Bo*, *JIN Ru-Long*, *WANG Lu*
(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

**Abstract**: Research shows that the role of body posture in emotion recognition is always underestimated. Aiming at the problems of low face resolution and low expression recognition accuracy in public space, an emotion recognition method based on facial expression and body posture is proposed. Firstly, the video data is preprocessed to obtain the input sequence of expression channel and posture channel; then, the emotional features of expression and posture are extracted by deep learning method; finally, fusion and classification are carried out in decision level. The emotion dataset (SCU-FABE) based on public space video is constructed. With this dataset, combined with posture emotion recognition data enhancement, the effective recognition of individual emotions in public space is realized. The experimental results show that the recognition rate of expression and posture is 94.698% and 88.024% respectively; the accuracy of fusion emotion recognition rate is 95.766%. The proposed method effectively integrates emotional information expressed by facial expression and body posture, which has good generalization ability and applicability in real scene video data.

**Keywords**: Deep learning; Emotion recognition; Decision-level fusion; Facial expression; Posture

**收稿日期**：2020-08-31
**基金项目**：国家自然科学基金(61871278)；四川省科技计划(2018HH0143)
**作者简介**：文虹茜(1997－)，女，云南楚雄人，硕士研究生，研究方向为深度学习. E-mail：736237443@qq.com
**通讯作者**：卿粼波. E-mail：qing_lb@scu.edu.cn

# 1 引　言

情绪的感知和表达在心理学和神经科学领域已经得到了广泛的研究,随着人工智能的不断发展,利用计算机进行情绪的分析也获得了人们的关注.能够感知和分析人类情绪和意图的计算机系统将在娱乐、医疗、教育和公共安全等领域发挥作用.例如,提高机器人情绪识别能力将丰富人机交互应用;情绪感知医疗辅助系统可以帮助评估焦虑和抑郁等精神障碍;在机场、地铁和公园等人流量大的场所进行情绪监测可以帮助识别潜在威胁,及时处理突发事件.

面部表情可以最直观地反映出人们的情绪状态和心理活动,是表达情绪的重要方式.目前基于视觉感知的人类情感的研究主要集中在面部.心理学家 Ekman[1] 研究不同文化之间的面部行为模式,定义了 6 类基本情绪(快乐、悲伤、厌恶、惊讶、愤怒和恐惧).传统的表情识别研究大多采用手工特征或浅层学习,随着应用环境转向具有挑战性的真实场景,神经网络越来越多地被用于特征提取,并取得了超前的识别精度.在表情识别中应用广泛的深度学习技术有卷积神经网络(Convolutional Neural Networks, CNN)、深度置信网络(Deep Belief Network, DBN)、递归神经网络(Recursive Neural Network, RNN)等[2].然而,有心理研究表明,面部表情本身可能包含误导性信息,特别是应用于互动和社交场景时.而通过观察身体姿势、动作、语调等不同的表现形式能提高对情绪状态的判断能力[3-4].此外,在真实环境中,距离、姿势、光照等因素会对面部情绪的识别产生很大影响,人脸分辨率不高,面部特征模糊,会降低面部表情识别率.

近年来,越来越多的情感神经科学研究表明,身体姿态在情感表达中与面部一样具有诊断性[5],姿态表现出来的倾斜方向、身体开放度和手臂、肩膀、头部位置等信息对情感状态的识别是有贡献的.通过连接到身体的传感设备可以感知人体位置和运动,获得的特征通常以骨骼的形式来进行情绪识别[6-7].然而,传感技术的使用存在诸多限制和差异,基于视觉的姿态情绪识别技术在图像、视频数据上的使用更加广泛.目前关于身体情绪表达的研究较少,大多使用人工提取特征的方法.但是在当今数据量越发巨大、数据越发复杂的情况下,手工设计和提取特征将耗费巨大的计算代价.

以前的人工特征或深度学习情绪识别工作使用单一的模式,如面部表情[8-12]、言语[13]、步态[6]以及生理信号[14]等.多模态情绪识别受到心理学研究的启发,情感的表达方式不是孤立存在的,这也有助于提高野外情绪识别的准确性[3].其中,面部表情和身体姿态的组合视觉渠道被认为是判断人类行为线索的重要渠道[15].有关融合表情及姿态的情绪识别文献很少,大多使用传统方法提取融合来自面部表情、身体姿态或者手势的线索.Gunes 等[15]基于轮廓和肤色跟踪头部和手部并提取了两百多个特征用于情绪识别,特征提取操作复杂,只使用了来自 4 个受试者的 27 个视频,数据量非常有限.Chen 等[16]使用运动历史图像(Motion History Image, MHI)方向梯度直方图(Histogram of Oriented Gradient, HOG)和图像方向梯度直方图的方法表示人脸和人体手势的局部运动信息和外观信息,提取的特征向量更加庞大.王晓华等[17]提出时空局部三值模式矩(TSLTPM),融合 3 维梯度方向直方图(3DHOG)特征描述纹理变化.姜明星等[18]使用时空局部三值方向角模式进行特征提取.Mittal 等[3]使用了静态的人脸和步态信息进行情绪识别,然而运动对于识别身体表达的情绪是十分重要的[5].神经网络的快速发展[19]使情感识别与分析领域也取得很大进步[20-21].然而由于缺乏大型的表情及姿态情绪数据集,表情及姿态融合情绪识别研究的潜力还待发掘.

本文针对公共空间个体人脸分辨率较低、面部特征模糊的问题,提出了融合表情及姿态的情绪识别方法.首先,对视频数据进行预处理获得表情通道和姿态通道的输入流;使用深度学习的方法实现表情和姿态情绪特征构建过程的自动化,减少计算复杂度;最后,在决策层进行融合和分类.通过有效融合表情和姿态在情绪识别中独特的优势,实现了公共空间个体情绪状态的有效识别.

# 2 融合表情及姿态的情绪识别

目前融合表情及姿态的情绪识别大多研究纯色背景实验室环境中采集的数据,人工构建和提取特征,多种特征提取技术的局限性在不断积累,降低了模型的泛化能力.而且使用手工特征将导致大量的计算开销,处理无约束情形下的大量数据会是巨大挑战.本文使用基于视觉的表情和姿态来扩展情绪识别的通道,提出基于深度学习的双通道情绪识别模型(如图 1).模型由数据预处理、特征提取和模式融合 3 个部分组成.为提供面部通道和姿态

通道的输入流，首先对原始数据进行预处理，包括面部检测、面部和身体视频序列尺寸处理. 针对表情进行空间流静态图像特征学习；对于姿态情绪，外观特征和运动特征都有重要作用，需要提取视频序列中有效的时空信息. CNN 网络具有很强的图像特征学习能力，不依赖人工经验；3DCNN 能同时学习时空特征，因此，本文采用两个网络分别对表情图像信息及姿态外观和运动信息进行建模. 最后，将两通道的输出加权融合并得到最终的分类结果.

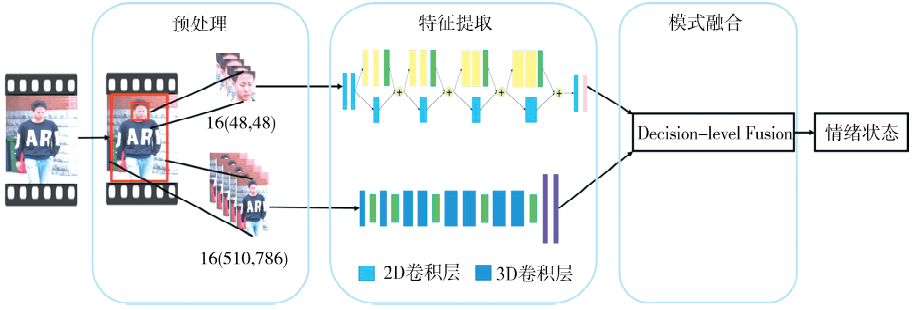


图 1　模型结构
Fig. 1　Architecture of model

## 2.1　数据预处理

数据预处理部分包括面部检测、面部和身体视频序列尺寸处理. 为了提供面部通道的输入流，本文使用多任务卷积神经网络 MTCNN[22] 进行面部检测. 将所有帧通过 MTCNN 得到面部图像，并调整为 48×48 像素. 双通道中的身体通道输入为视频序列，所有视频帧尺寸统一调整为 510×786 像素.

## 2.2　特征提取

2.2.1　面部通道　为获得面部表情信息，使用深度可分离卷积神经网络 Mini-Xception[23] 进行特征提取. Mini-Xception 的网络模型来源于 Xception 架构，深度可分离卷积能更加有效地利用模型参数，残差连接模块能加快收敛过程，结构如图 2 所示. 通过 Mini-Xception 能自动提取面部输入的有效特征，为与身体通道的融合做准备. 训练阶段学习率设置为 0.1，批量大小 32，使用早停法防止过拟合.

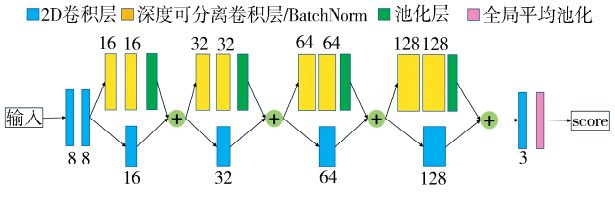


图 2　Mini-Xception 结构
Fig. 2　Architecture of Mini-Xception

2.2.2　姿态通道　为了获得姿态情绪信息，使用 C3D 网络[24] 进行特征提取. 研究表明，外观和运动信息都对从身体表达中感知情绪起重要作用. 同时，对于视频序列，有效的时空信息也很关键. C3D 能简单高效地学习时空特征，关注外观和运动信息，适合用于身体姿态情绪特征的提取. C3D 网络结构示意图如图 3，训练阶段初始学习率为 0.001，批量大小 10.

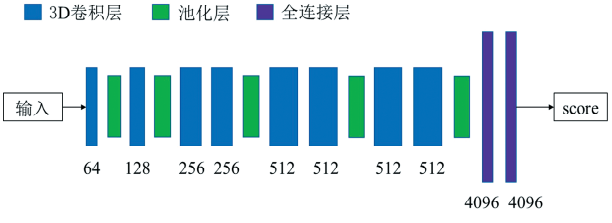


图 3　C3D 结构
Fig. 3　Architecture of C3D

## 2.3　通道融合

面部通道和身体通道获得的特征信息各有优势，将两个通道融合进行分类. 采用加权和的决策层融合：使用神经网络学习特征后，在全连接层后获得类别的后验概率，将面部和身体两个通道输出的后验概率求加权和. 因为面部表情是主要模式，因此面部通道和身体通道的权重分别设置为 0.7 和 0.3.

# 3　公共空间个体情绪数据集构建

目前利用表情及姿态进行情绪识别的研究较少，可以直接用于训练的数据集也十分匮乏. Gunes 等人[15] 在实验室中收集了包含面部和上身的情绪数据集 FABO，此后相关研究大多基于此数据集开展. 然而 FABO 标注不全，23 个受试者中只

有 16 个具有标注;样本数量很少且情感类别不均匀,利用深度学习方法训练时容易出现过拟合现象,因此无法利用 FABO 开展本文研究. Bänziger 等[25]创建了日内瓦多模态情感刻画(GEMEP)数据集,数据集包含了来自实验室的 10 个受试者的面部和身体的视频及语音. 然而 GEMEP 并未公开发布,无法用于本文个体情绪的研究.

通过定点拍摄、网上搜集和真人表演 3 种方式建立公共空间个体情绪数据集 SCU-FABE. 首先,利用 KCF 跟踪算法[26]对视频中的行人进行跟踪和保存,KCF 算法具有准确度高、运算速度快的双重优势,适用于少量行人目标的跟踪. 然后,剔除不合格的个体序列再进行情绪标注. 情感计算领域使用比较广泛的模型有离散型和连续型. 连续型并不适用于城市公共空间中个体的情绪划分,因为在公共空间中人流密度大,对视频中的每一个人进行精细化的情绪分析耗时耗力. 相对于判断情感程度,识别个体情绪的正负性更为首要. Russell[27]提出的 Arousal-Valence 模型中价效(Valence)表征了情感的正负性. SCU-FABE 主要从价效出发,将情绪划分为消极、中性和积极三类,邀请 10 名志愿者(5 名男性和 5 名女性)进行手动标注. 实验中总共使用公共空间个体情绪数据序列 993 个,每个序列的长度为 20 帧到 100 帧不等. 其中 Negative 类包含 324 个序列、Neutral 类包含 315 个序列、Positive 类包含 354 个序列,按照接近 1∶1 的比例划分训练集和测试集. 图 4 为表达序列示例.


图 4　数据集表达序列示例
(a)"消极"序列;(b)"积极"序列;(c)"中性"序列
Fig. 4　Samples of dataset
(a) negative; (b) positive; (c) neutral

## 4　实验结果与分析

### 4.1　实验设置

本文在基于 Python 的深度学习框架 TensorFlow 环境下进行实验. 实验环境为: Ubuntu 18.04, NVIDIA Tesla K80 GPU. 为评估本文提出的融合表情及姿态的情绪识别性能,进行如下实验:(1) 数据增强实验,探究针对姿态数据情绪识别的数据增强方法;(2) 面部情绪识别实验和姿态情绪识别实验,作为单模式情绪识别对照组,与融合的情绪识别结果进行对比分析;(3) 融合情绪识别实验,验证融合表情及姿态的情绪识别方法的有效性.

### 4.2　数据增强实验

神经网络需要大量的数据训练才能获得更好的性能. 对于面部数据,已验证过可靠性和有效性的数据增强方法有很多,最常用的方法包括旋转、平移、翻转、随机裁剪和随机加入噪声等等,可以很好地扩充数据集,增强模型的泛化能力. 然而对于姿态数据,使用常见的扩充数据的操作是否会破坏身体姿态序列在情绪识别方面潜在的重要特征是一个需要探究的问题.

为了更有效地扩充数据、完成情绪识别目标,针对身体姿态数据分别使用原始数据、颜色处理数据、旋转处理数据以及镜像处理数据进行扩充. 方案基于以下假设:未处理的原始数据不会丢失情绪识别相关线索. 在唯一变量为输入数据的情况下训练和测试,以原始数据的测试结果为阈值,已处理数据的测试结果低于此阈值则判断为有破坏相关线索的可能. 数据处理对比图如图 5 所示.

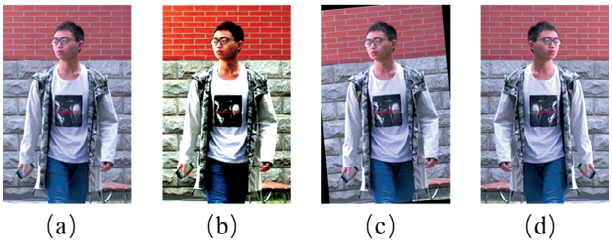


图 5　数据处理对比图
(a)原始图像;(b)颜色处理;(c)旋转处理;(d)镜像处理
Fig. 5　Samples of data processing
(a) Original image; (b) Color processing;
(c) Rotation processing; (d) Mirror image

使用测试集进行测试,因为样本数量比较均衡,以 10 次测试结果的平均识别率为评价指标. 实验结果表明,有关外观和运动的信息都对情绪感知

有作用，颜色处理和旋转处理加强了潜在特征，识别率更高；镜像处理破坏了潜在特征，识别率更低. 最终训练使用 10%自动对比度和逆时针旋转 5°的方法进行处理，数据量扩充为原来的 3 倍. 分别使用原始数据和扩充后的数据进行训练，测试结果对比如表 1. 实验结果表明，使用颜色处理和旋转处理的方法进行数据增强效果比较明显，识别率提高了 5.927%.

**表 1　数据增强实验结果**

Tab. 1　The experiment results of data enhancement

| 方案 | 原始数据 | 颜色处理 | 旋转处理 | 镜像处理 | 数据增强 |
|---|---|---|---|---|---|
| 平均识别率/% | 82.097 | 83.407 | 85.504 | 79.879 | 88.024 |

### 4.3　单模式情绪识别实验

为了验证单独的面部和姿态对情绪识别的作用以及作为融合模式的双通道产生的贡献，进行单模式情绪识别对照实验. 使用经过预处理和数据增强的训练集进行训练，面部序列和姿态序列是相互对应的. 使用测试集进行 10 次测试，采用平均识别率作为评价指标.

从表 2 实验结果可知，面部对于情绪识别有重要意义，平均识别率为 94.698%，从表 3 混淆矩阵可知，通过面部感知“消极”情绪的效果最差，容易误判为“中性”情绪. 身体姿态在情绪表达中具有诊断性，能自发揭示一些情绪线索，平均识别率为 88.024%. 从表 4 混淆矩阵可知，通过身体姿态感知“积极”情绪的效果最差.

**表 2　情绪识别实验结果**

Tab. 2　Emotion recognition results

| 方案 | 表情 | 姿态 | 融合 |
|---|---|---|---|
| 平均识别率/% | 94.698 | 88.024 | 95.766 |

**表 3　面部情绪识别混淆矩阵**

Tab. 3　Facial emotion recognition confusion matrix

| 情绪状态 | 消极/% | 中性/% | 积极/% |
|---|---|---|---|
| 消极 | 92.5 | 6.79 | 0.62 |
| 中性 | 2.55 | 94.90 | 2.55 |
| 积极 | 1.13 | 2.82 | 96.00 |

**表 4　姿态情绪识别混淆矩阵**

Tab. 4　Posture emotion recognition confusion matrix

| 情绪状态 | 消极/% | 中性/% | 积极/% |
|---|---|---|---|
| 消极 | 90.70 | 9.26 | 0.00 |
| 中性 | 8.92 | 90.40 | 0.63 |
| 积极 | 7.34 | 10.7 | 81.90 |

### 4.4　融合情绪识别实验

如表 5 所示，融合情绪识别实验验证了通过表情和姿态进行情绪识别的有效性，平均识别率达到 95.766%，高于单独的面部情绪识别和姿态情绪识别. 通过对比单模式和融合情绪识别混淆矩阵可以更加直观的看出融合模式的优势：当两个通道融合时，面部感知“消极”情绪的局限和身体感知“积极”情绪的局限得到互补改进，“中性”情绪的识别率提高，从而获得整体判决正确率的提高. 说明面部表情和身体姿态都对情绪识别有所贡献，并且表达的信息可有效地互补，结合面部表情和身体姿态能提高识别情绪状态的能力和可靠性.

**表 5　融合情绪识别混淆矩阵**

Tab. 5　Fusion emotion recognition confusion matrix

| 情绪状态 | 消极/% | 中性/% | 积极/% |
|---|---|---|---|
| 消极 | 95.00 | 4.32 | 0.62 |
| 中性 | 1.27 | 98.10 | 0.63 |
| 积极 | 0.56 | 3.95 | 95.40 |

## 5　结　论

本文设计了一种融合表情及姿态的情绪识别方法，使用两个通道提取面部和身体与情绪有关的信息，在决策层进行融合和分类. 实验结果表明，对于大量真实场景视频数据，本文方法具有良好的泛化能力和适用性；表情和姿态表达的情感信息具有较好的互补作用，结合使用能提高情绪识别可靠性. 对于身体姿态情绪识别，使用深度学习的方法自动提取特征取得了很好的效果，表明身体姿态情绪识别从使用几何表示的简单静态或动态特征转向深度学习表征具有很大的潜力.

本文的研究针对公共空间个体情绪识别，而公共空间中多人群组普遍存在，表达的情绪之间存在相关性，对于人群整体情绪的计算也十分有意义. 研究公共空间中多尺度情绪识别是下一步所要做的工作.

**参考文献：**

[1]　Ekman P. Facial expression and emotion [J]. Am Psychol, 1993, 48: 384.

[2]　Li S, Deng W. Deep facial expression recognition: a survey [J]. IEEE TAffect Comput, 2020, 99: 1.

[3]　Mittal T, Guhan P, Bhattacharya U, *et al*. EmotiCon: context-Aware multimodal emotion recognition using Frege's principle [C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR). Seattle: IEEE, 2020.

[4] Sun B, Cao S, He J, *et al*. Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy [J]. Neural Networks, 2017, 105: 36.

[5] Luo Y, Ye J, Adams R B, *et al*. ARBEE: Towards automated recognition of bodily expression of emotion in the wild [J]. Int J Comput Vision, 2020, 128: 1.

[6] Randhavane T, Bhattacharya U, Kapsaskis K, *et al*. Identifying emotions from walking using affective and deep features [J]. arXiv preprint arXiv: 2019, 1906: 11884.

[7] Saha S, Datta S, Konar A, *et al*. A study on emotion recognition from body gestures using Kinect sensor [C]// Proceedings of the 2014 International Conference on Communication and Signal Processing. Melmaruvathur: IEEE, 2014.

[8] Kollias D, Zafeiriou S P. Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the OMG in-the-wild dataset [J]. IEEE T Affect Comput, 2020, 99: 1.

[9] Yang H, Ciftci U, Yin L, *et al*. Facial expression recognition by de-expression residue learning [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018.

[10] 李婷婷, 胡玉龙, 魏枫林. 基于GAN改进的人脸表情识别算法及应用[J]. 吉林大学学报: 理学版, 2020, 58: 605.

[11] 马玉环, 张瑞军, 武晨, 等. 深度残差网络和LSTM结合的图像序列表情识别[J]. 重庆邮电大学学报: 自然科学版, 2020, 32: 874.

[12] 何明. 一种深度自编码器面部表情识别新方法[J]. 西南师范大学学报: 自然科学版, 2019, 44: 81.

[13] Zhao J, Mao X, Chen L, *et al*. Speech emotion recognition using deep 1D & 2D CNN LSTM networks [J]. Biomed Signal Proces, 2019, 47: 312.

[14] Hassan M M, Alam M G, Uddin M Z, *et al*. Human emotion recognition using deep belief network architecture [J]. Inform Fusion, 2019, 51: 10.

[15] Gunes H, Piccardi M. Bi-modal emotion recognition from expressive face and body gestures [J]. J Netw Comput Appl, 2007, 30: 1334.

[16] Chen S, Tian Y, Liu Q, *et al*. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features [J]. Image Vis Comput, 2013, 31: 175.

[17] 王晓华, 侯登永, 胡敏, 等. 复合时空特征的双模态情感识别 [J]. 中国图象图形学报, 2017, 22: 39.

[18] 姜明星, 胡敏, 王晓华, 等. 视频序列中表情和姿态的双模态情感识别 [J]. 激光与光电子学进展, 2018, 630: 167.

[19] 徐富勇, 余谅, 盛钟松. 基于深度学习的任意形状场景文字识别 [J]. 四川大学学报: 自然科学版, 2020, 57: 255.

[20] 赵容梅, 熊熙, 琚生根, 等. 基于混合神经网络的中文隐式情感分析 [J]. 四川大学学报: 自然科学版, 2020, 57: 264.

[21] 刘广峰, 黄贤英, 刘小洋, 等. 基于主题注意力层次记忆网络的文档情感建模 [J]. 四川大学学报: 自然科学版, 2019, 56: 833.

[22] Zhang K, Zhang Z, Li Z, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Proc Let, 2016, 23: 1499.

[23] Arriaga O Arriaga O, Valdenegro-Toro M, *et al*. Real-time convolutional neural networks for emotion and gender classification [J]. arXiv preprint arXiv: 2017, 1710: 07557.

[24] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015.

[25] Bänziger T, Mortillaro M, Scherer K R. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception [J]. Emotion, 2012, 12: 1161.

[26] Henriques J F, Caseiro R, Martins P, *et al*. High-speed tracking with kernelized correlation filters [J]. IEEE T Pattern Anal, 2014, 37: 583.

[27] Russell J A. A circumplex model of affect [J]. J Pers Soc Psychol, 1980, 39: 1161.

**引用本文格式:**

中　文: 文虹茜, 卿粼波, 晋儒龙, 等. 基于表情及姿态融合的情绪识别[J]. 四川大学学报: 自然科学版, 2021, 58: 043002.

英　文: Wen H Q, Qing L B, Jin R L, *et al*. Emotion recognition based on fusion of expression and posture [J]. J Sichuan Univ: Nat Sci Ed, 2021, 58: 043002.