Computer Science and Application 计算机科学与应用, 2023, 13(6), 1257-1263
Published Online June 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.136123

Hans汉斯

基于自注意力机制与特征融合的课堂学生表情识别模型

陈夫叶，牛连强

沈阳工业大学软件学院，辽宁 沈阳

收稿日期：2023年5月21日；录用日期：2023年6月20日；发布日期：2023年6月27日

摘　要

为解决通常课堂场景下学生人脸表情识别的遮挡问题，通过部分分割和随机遮挡策略将原图分割成多路人脸图像，采用相同的残差网络提取特征，借助自注意力机制为多路网络分配不同权重，再对损失函数进行约束以限制遮挡支路权重始终小于眼部支路权重，进而得到加权后的支路特征并通过特征融合形成全局特征。在公开数据集FERplus上实验表明，模型能够大幅度提升人脸表情识别的准确率，有效缓解复杂场景下因遮挡造成的信息损失问题。

关键词

特征融合，自注意力机制，人脸表情识别

Classroom Student Expression Recognition Model Based on Self Attention Mechanism and Feature Fusion

Fuye Chen, Lianqiang Niu

School of Software, Shenyang University of Technology, Shenyang Liaoning

Received: May 21st, 2023; accepted: Jun. 20th, 2023; published: Jun. 27th, 2023

Abstract

In order to solve the occlusion problem of students' facial expression recognition in common classroom scenes, the original image is divided into multiple face images through partial segmentation and random occlusion strategies, the same residual network is used to extract features, the

文章引用: 陈夫叶, 牛连强. 基于自注意力机制与特征融合的课堂学生表情识别模型[J]. 计算机科学与应用, 2023, 13(6): 1257-1263. DOI: 10.12677/csa.2023.136123

self attention mechanism is used to assign different weights to the multiple networks, and then the loss function is constrained to limit the weight of the occlusion branch to always be less than the weight of the eye branch. Then, the weighted branch features are obtained and global features are formed through feature fusion. Experiments on the public dataset FERplus have shown that the model can significantly improve the accuracy of facial expression recognition and effectively alleviate the problem of information loss caused by occlusion in complex scenes.

Keywords

Feature Fusion, Self Attention Mechanism, Facial Expression Recognition

1. 引言

人类主要是通过面部表情这一途径来传达自身情感信息的。早在上世纪 70 年代就出现了表情编码系统的概念，人脸表情被分为惊讶、生气、高兴、厌恶、悲伤、害怕 6 类。部分表情识别算法已能够实现较为准确的表情分类，如 Zhang H 等[1]提出的弱监督局部全局关系网络，Marrero 等[2]提出的带有 Attention 机制的基于 CNN 的网络架构 FERAtt，Li 等[3]提出的基于注意力机制的卷积神经网络 ACNN，Fard 等[4]提出了一种自适应相关损失，以指导网络生成类内样本相关性高、类间样本相关性较低的嵌入式特征向量，改善嵌入式特征向量之间对类类别的区分。张海峰[5]等提出的基于多特征融合的人脸表情识别模型。这些模型的底层依赖 AlexNet [6]、VGGNet [7]、GoogleNet [8]、ResNet [9]等网络。

学生群体课堂学习情况分析是一种特殊场景，此时的面部表情是个体听课专注度的反映，但信息的准确获取所受影响因素较多，如头部姿态、背景信息、面部遮挡等。

为了消除复杂背景影响，Lee 等[10]设计了两个支路模型，各支路分别用于正常提取人脸面部特征和聚焦面部表情之外的背景信息，最后进行融合。Liu 等[11]提出双分支多特征学习网络，有效地区分局部面部特征的细微差异。Acharya 等[12]使用流形网络结构进行协方差合并，再使用二阶统计量捕捉面部特征扭曲，以更好地捕获部分变形的区域面部特征。Zhou 等[13]利用 Attention 机制和双线性池化进行多模态的表情特征融合，以使模型专注于面部的重要部位，提升面部表情识别的准确性。Wang 等[14]运用对抗学习思想消除身体姿态变化干扰，得到单纯的人脸表情特征，提高了表情识别的鲁棒性。Zhong 等[15]利用图结构的人脸表情表示和双向循环神经网络进行特征提取，有效去除了冗余信息，减少了干扰和训练开销。

与普通的人群密集型场景不同，学生课堂听课时的面部遮挡主要集中在左脸、嘴巴、右脸等下半张脸。为此，本文采取了如下的解决方法：

1) 引入自注意力机制学习局部特征并与整体特征融合，使模型更关注表情有效区域，过滤无效信息，获取表情细节特征。

2) 限制遮挡支路权重始终小于眼部支路权重，并通过阈值限定眼部区域权重在合理区间从而进一步弥补课堂场景下面部遮挡带来的信息损失，提高模型准确率。

2. 基于自注意力机制与特征融合的表情识别模型

模型共分为提取人脸表情特征、自注意力权重分配、特征融合和表情分类 4 部分，如图 1 所示。

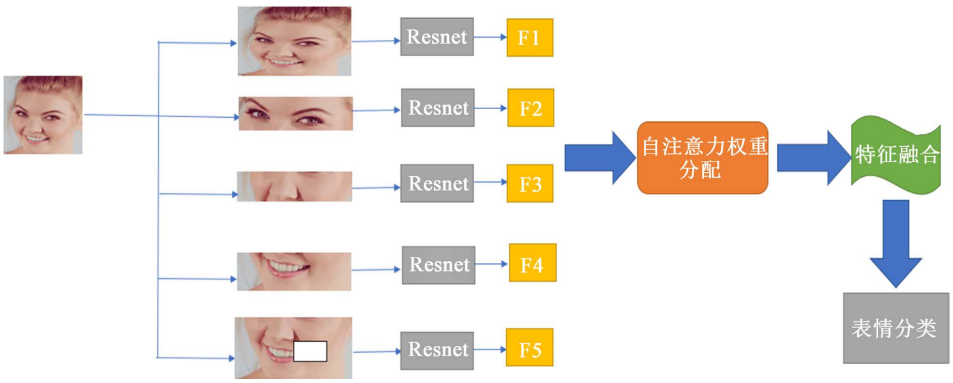

Figure 1. Model framework
图 1. 模型框架

在提取人脸表情特征模块，先对课堂教学视频帧中的原人脸图像进行裁剪和遮挡，将裁剪和遮挡后的人脸图像与原人脸图像分成五路，每路都通过 Resnet 残差网络提取局部特征。

在自注意力权重分配模块分别对从每个支路学习到的学生人脸表情特征分配权重，自注意力机制会为从各支路学习到的表情特征分配权重，然后再通过约束性损失函数约束分配到的权重使得眼部特征分配到最大权重。

在特征融合模块将从注意力分配特征权重模块得到的各支路权重和学生人脸表情特征进行融合得到学生人脸表情的全局特征，学生人脸表情全局特征由各个支路加权注意力特征之和得到。

最后在表情分类模块将得到的数据使用 SoftMax 分类器分类学生人脸表情。

3. 表情特征学习

3.1. 特征提取

在人脸图像中，嘴部、眼部和鼻子为识别提供了局部特征，可供于融合成全脸图。因此，图 1 设计了 5 条支路，分别提取整体特征、眼部特征、嘴部特征、鼻子部位特征和下半张脸遮挡特征。特别地，将下半张脸随机遮挡作为特征有利于解决因课堂脸部遮挡问题而产生的表情识别准确率低问题。

模型先根据人脸检测得到的人脸关键点坐标信息裁剪各支路图像，并使用数据增强方法进行下半张脸随机遮挡。因为裁剪得到的不同部位图像尺寸不同，需将各支路图像缩放为相同尺寸，并对下半张脸中包含的鼻子和嘴部进行像素值归一化处理，最后送进支路网络。

各支路网络由相同的 ResNet 残差网络组成，其后端连接全连接层并使用激活函数来获取各支路提取到的特征所占权重，同时引用约束性损失函数约束遮挡支路的特征权重，使其小于眼部特征权重最后融合整体人脸表情特征和局部人脸表情特征得到表情特征图。具体模型图参见图 2。

3.2. 自注意力权重分配

公图像输入部分输入学生人脸图像得到人脸复制图 X_0。部位细分部分由 X_0 输出 X_1~X_4，分别对应眼部、鼻子、嘴部和随机遮挡的下半张脸。对应地，在特征提取部分得到输出 F_0~F_4，而注意力权重分配部分依据式(1)计算 F_0~F_4 的权重 μ_i。

$$\mu_i = f\left(W_2 * R\left(W_1 * F_i\right)\right), 0 \le i \le 4 \tag{1}$$

其中，W_1 和 W_2 为全连接层权重，R 为 ReLU 激活函数，f 为 Sigmoid 激活函数。由各支路 ResNet 残差网络学习到的学生人脸表情特征 F_i 与注意力权重 μ_i 的加权和构成了式(2)所示的全局人脸表情特征 F_m。

$$F_m = \frac{\sum_{i=0}^{4} \mu_i * F_i}{\sum_{i=0}^{4} \mu_i} \tag{2}$$

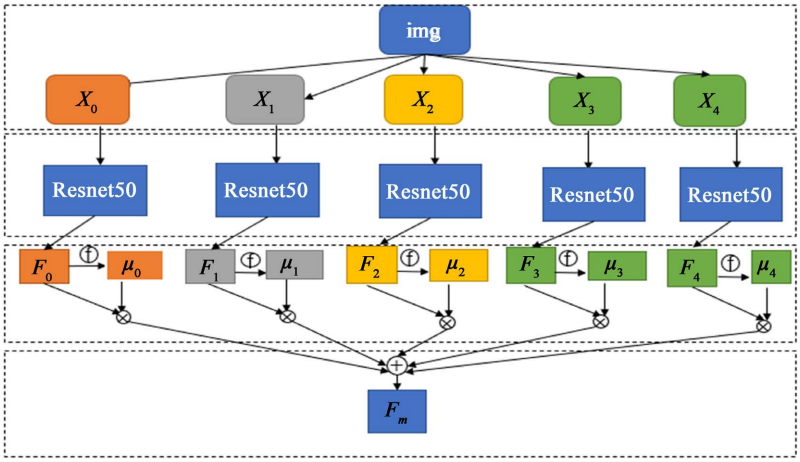

Figure 2. Student expression feature learning model diagram
图 2. 学生表情特征学习模型图

3.3. 约束性损失与分类损失函数

由于课堂场景下遮挡区域主要集中在下半张脸，应使模型减少对遮挡区域的关注。因此，采用式(3)所示的约束性损失函数，以限制遮挡支路权重小于眼部支路权重。

$$L_1 = \max\left(0, margin - \left(\mu_e - \mu_m\right)\right) \tag{3}$$

其中，μ_m 为遮挡支路权重值，μ_e 为眼部支路权重，$margin$ 为阈值初始值设为 0.03。当眼部支路的权重值与遮挡支路的权重值之间的差距超过了阈值 $margin$ 时，就会产生损失，目的是让模型更加关注眼部支路的特征。具体来说当眼部支路的权重值 μ_e 大于遮挡支路的权重值 μ_m 的加上 $margin$ 时，损失函数的值为 0，说明网络更加关注眼部特征，对于眼部的识别和定位更加准确，不会对网络进行惩罚。当眼部支路的权重值 μ_e 小于遮挡支路的权重值的加上 $margin$ 时，损失函数的值会大于 0，说明网络更加关注遮挡区域特征，这是会对网络进行惩罚，从而引导网络更加关注眼部特征。

函数 L_1 鼓励模型对每个样本预测其真实标签的概率尽可能大，并对预测概率和其他概率之间的差距进行限制。

分类损失采用式(4)所示的交叉熵函数。

$$L_2 = y \log y' - (1-y)\log(1-y') \tag{4}$$

其中，y 为真实人脸表情标签，y' 为预测人脸表情标签。

最后，将约束性损失与分类损失求和作为总损失函数，参见式(5)。

$$L = L_1 + L_2 \tag{5}$$

4. 实验与结果分析

4.1. 数据集与参数设置

本文使用公开数据集 FERplus [13]进行验证，它是谷歌搜索引擎收集的大规模真实人脸表情数据集，包括 28,709 张训练图像，3589 张验证图像和 3589 张测试图像。

残差网络采用 ResNet50；训练数据集学习率初始化为 0.01；Epoch 迭代周期为 30，学习率每迭代一个周期更新为当前的 0.1 倍。

4.2. 对比分析

表 1 显示了本文模型与 PLD (Probabilistic Label Drawing) [16]、Deep-emotions [17]和 Efficient-Net [18]在测试集上的准确率。

Table 1. Accuracy of facial expression recognition using several models
表 1. 几种模型的表情识别准确率

模型	准确率
Deep-emotion	65.68%
Efficient-Net	83.6%
PLD	85.1%
本模型	87.91%

在模型中，约束性损失函数的阈值 *margin* 设置为 0.03。可以发现，由于融合了局部与整体表情特征，重点关注眼部这一不常被遮挡的人脸区域，本文模型更准确地获得了特殊场景下的表情特征，进而提高了表情识别的准确率。

鉴于阈值 *margin* 的重要性，图 3 比较了其值分别取 0.01、0.03、0.05、0.07 和 0.1 时对识别准确率的影响。

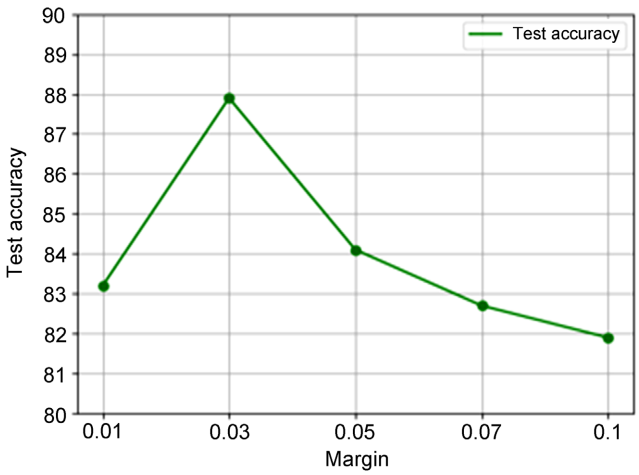

Figure 3. Student expression feature learning model diagram
图 3. 学生表情特征学习模型图

当 *margin* 值过大时，模型会过于关注眼部区域，使其他区域的特征影响降低而导致识别准确率下降。当 *margin* 值过小时，模型对所有特征区域的关注程度相对均衡，无法将权重偏向于不常被遮挡的眼部区域。

5. 结论

为了解决课堂教学场景下学生人脸表情识别问题，构建了利用局部特征和整体特征的融合，并使用

约束性损失函数对遮挡支路进行限制，重点关注不常被遮挡的眼部区域的识别模型在公开数据集上较现有模型得到了更高的表情识别准确率。模型表明，对于一些有明显约束的场景，通过布局和整体的特征融合及对损失函数进行约束可以有效缓解因某些信息缺失而造成的识别准确率下降问题。结合对眼部动态的判别，可以利用该模型进一步衡量学生课堂学习的专注度。

参考文献

[1] Zhang, H., Su, W., Yu, J., *et al*. (2020) Weakly Supervised Local-Global Relation Network for facial Expression Recognition. *Proceedings of Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence* (*IJCAI-PRICAI*-2020), Yokohama, 7-11 July 2020, 1040-1046. https://doi.org/10.24963/ijcai.2020/145

[2] Fernandez, P.D.M., Peña, F.A.G., Ren, T.I. and Cunha, A. (2019) Feratt: Facial Expression Recognition with Attention Net. *Proceedings of the* 2019 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, 16-17 June 2019, 837-846. https://doi.org/10.1109/CVPRW.2019.00112

[3] Li, Y., Zeng, J., Shan, S. and Chen, X. (2018) Occlusion Aware Facial Expression Recognition Using CNN with Attention Mechanism. *IEEE Transactions on Image Processing*, **28**, 2439-2450. https://doi.org/10.1109/TIP.2018.2886767

[4] Fard, Ali P. and Mahoor, M.H. (2022) Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild. *IEEE Access*, **10**, 26756-26768. https://doi.org/10.1109/ACCESS.2022.3156598

[5] 张海峰. 基于多特征融合的人脸表情识别研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2020.

[6] Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90.

[7] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of* 3*rd International Conference on Learning Representations* (*ICLR* 2015), San Diego, 7-9 May 2015, 1-14.

[8] Szegedy, C., Wei, L., Jia, Y., *et al*. (IEEE) Going Deeper with Convolutions. *Proceedings of the* 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[9] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the* 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[10] Lee, J., Kim, S., Kim, S., *et al*. (2019) Context-Aware Emotion Recognition Networks. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Seoul, 27 October-2 November 2019, 10143-10152. https://doi.org/10.1109/ICCV.2019.01024

[11] Liu, X., Guo, Z., Yuan, B. and Guo, H. (2022) Robust Facial Expression Recognition Based on Dual Branch Multi-feature Learning. *Proceedings of* 2022 7*th International Conference on Image*, *Vision and Computing* (*ICIVC*), Xi'an, 26-28 July 2022, 1-6. https://doi.org/10.1109/ICIVC55077.2022.9886565

[12] Acharya, D., Huang, Z., Paudel, D.P. and Van Cool, L. (2018) Covariance Pooling for Facial Expression Recognition. *Proceedings of* 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, 18-22 June 2018, 367-374. https://doi.org/10.1109/CVPRW.2018.00077

[13] Zhou, H., Meng, D., Zhang, Y., *et al*. (2019) Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. *Proceedings of* 2019 *International Conference on Multimodal Interaction*, Suzhou, 14-18 October 2019, 562-566. https://doi.org/10.1145/3340555.3355713

[14] Wang, C., Wang, S. and Liang, G. (2019) Identity- and Pose-Robust Facial Expression Recognition through Adversarial Feature Learning. *Proceedings of the* 27*th ACM International Conference on Multimedia Interaction*, Nice, 21-25 October 2019, 238-246. https://doi.org/10.1145/3343031.3350872

[15] Zhong, L., Bai, C., Li, J., *et al*. (2019) A Graph-Structured Representation with BRNN for Static-based Facial Expression Recognition. *Proceedings of* 2019 14*th IEEE International Conference on Automatic Face & Gesture Recognition*, Lille, 14-18 May 2019, 1-5. https://doi.org/10.1109/FG.2019.8756615

[16] Barsoum, E., Zhang, C., Ferrer, C.C., *et al*. (2016) Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. *Proceedings of the* 2016 *International Conference on Multimodal Interaction*, Tokyo, 12-16 November 2016, 279-283. https://doi.org/10.1145/2993148.2993165

[17] Abdolrashidi, A. (2021) Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors*, **21**, Article 3046. https://doi.org/10.3390/s21093046

[18] Siqueira, H., Magg, S. and Wermter, S. (2020) Efficient Facial Feature Learning with Wide Ensemble-Based Convolutional Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 5800-5809. https://doi.org/10.1609/aaai.v34i04.6037