基于注意力机制和混合网络的小群体情绪识别

季欣欣，邵 洁，钱勇生

（上海电力大学 电子与信息工程学院，上海 200090）

摘 要：针对自然状态下小群体图像的情绪分类，提出基于面部、场景和骨架 3 种视觉线索的混合深度网络，分别利用 3 类卷积神经网络（convolutional neural networks，CNN）分支独立学习，通过决策融合获得最终的情绪分类。其中面部 CNN 通过注意力机制学习不同人脸的权重，获得整张图片关于人脸的特征表示，利用 large-margin softmax（L-softmax）损失函数进行判别性学习；使用先进的姿势估计方法 OpenPose 获得图像中所有人体骨架，作为基于骨架卷积神经网络的输入。考虑图片的场景信息，将整张图片作为基于场景 CNN 的输入。实验结果表明，改进模型对自然状态下 3 种类型的小群体情绪识别鲁棒，取得了较高的准确率。

关键词：小群体情绪识别；场景理解；混合网络；注意力机制；大边缘损失函数；压缩和奖惩网络模块

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2020）06-1683-06

doi：10.16208/j.issn1000-7024.2020.06.030

Group emotion recognition based on attention mechanism and hybrid network

JI Xin-xin, SHAO Jie, QIAN Yong-sheng

(College of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 200090, China)

Abstract: Aiming at the emotional classification of small group images in natural state, a hybrid depth network based on three visual cues of face, scene and skeleton was proposed. The three types of convolutional neural networks (CNN) were used to learn independently and decision fusion was implemented to gain the final emotional classification. The facial CNN learnt the weights of different faces through the attention mechanism, obtained the feature representation of the whole picture about the face, and used the large-margin softmax (L-softmax) loss for discriminative learning. The advanced pose estimation method OpenPose was used to obtain all human skeletons in the image as input based on the skeleton convolutional neural network. The scene information of the picture was considered, and the whole picture was taken as the input based on the scene CNN. Experimental results show that the improved model is robust to three types of small group emotion recognition under natural conditions, and it has achieved high accuracy.

Key words: group emotion recognition; scene understanding; hybrid network; attention mechanism; large-margin softmax; squeeze-and-excitation block

0 引 言

智能情感分析研究已经走过了漫长的道路，但传统上一直关注场景中的单一个体，而不是群体[1]。群体可分为大小群体，大群体如街道的人流，此时人与人之间并没有情感的交流和统一的情绪，本文是对多位个体间有情感交流的小群体进行情绪识别，下文提及的群体均指小群体。由于面部遮挡、光照变化、头部姿势变化，各种室内和室外环境不同以及由于相机距离不同而导致低分辨率的面部图像，因此群体情绪识别问题具有挑战性。

目前，针对群体情绪识别已有许多研究方法。Dhall 等[2]介绍了 AFEW 数据库和群体情绪识别框架，包括使用面部动作单元提取面部特征，在对齐的面上提取低级特征，使用 GIST 和 CENTRIST 描述符提取场景特征并使用多核学习融合，但是他们提出的方法依赖于 LBQ 和 PHOG 特征和 CENTRIST，其捕获面部表示和场景表示是有限的。Y. Qiao[3]提出将基于面部和整张图像上的卷积神经网络（CNN）单独训练，并融合以得到分类结果。然而群体情绪

收稿日期：2019-04-01；修订日期：2019-05-21

基金项目：国家自然科学基金青年科学基金项目（61302151、61401268）；上海市自然科学基金项目（15ZR1418400）

作者简介：季欣欣（1994-），男，硕士研究生，研究方向为计算机视觉、深度学习；邵洁（1981-），女，博士研究生，副教授，研究方向为计算机视觉、图像处理；钱勇生（1992-），男，硕士研究生，研究方向为计算机视觉、深度学习。E-mail：737787764@qq.com

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

计算为群体成员的幸福水平的平均值，忽略了特殊个体信息（例如脸部的遮挡水平和哭笑的脸），因此对于群体情绪识别仍有待提高。

在本文中提出通过建立混合网络来解决这一问题，该网络在面部、场景和骨架上单独训练 3 个卷积神经网络（CNN）分支，然后通过决策融合以获得最终的情绪分类。其中一个模型是基于人脸面部特征来训练，并使用注意力机制学习不同人脸的权重，获得整张图片关于人脸的特征表示。通过对比实验，验证了本文方法的有效性，并获得较高的准确率。

1 群体情绪识别架构

本文的系统框架如图 1 所示。首先对检测到的人脸做对齐相似变换，作为面部 CNN 的输入，并通过注意力机制学习图像中不同人脸的权重，获得整张图片关于人脸的特征表示。其次使用 OpenPose 获得图像中人体的骨架，作为骨架 CNN 的输入。同时考虑了图片的场景信息，将整张图片作为场景 CNN 的输入。3 种类型的 CNN 都训练了多个模型，然后对选取的模型执行决策融合以学习最佳组合。

1.1 面部 CNN

群体图像中人脸所描绘的表情传达了充分的情感信息，在情绪识别中起着至关重要的作用，因此建立面部情感 CNN 来进行群体情绪识别。本文使用 ResNet18[4] 模型，模型的输入为对齐的人脸图像。为了减轻过拟合现象并增强模型泛化能力，使用 CASIA-Webface 数据集对其进行预训练，然后使用 L-softmax 损失在 EmotiW 训练数据集[5] 中进行微调。下面介绍本文使用 L-Softmax 损失函数和注意力机制。

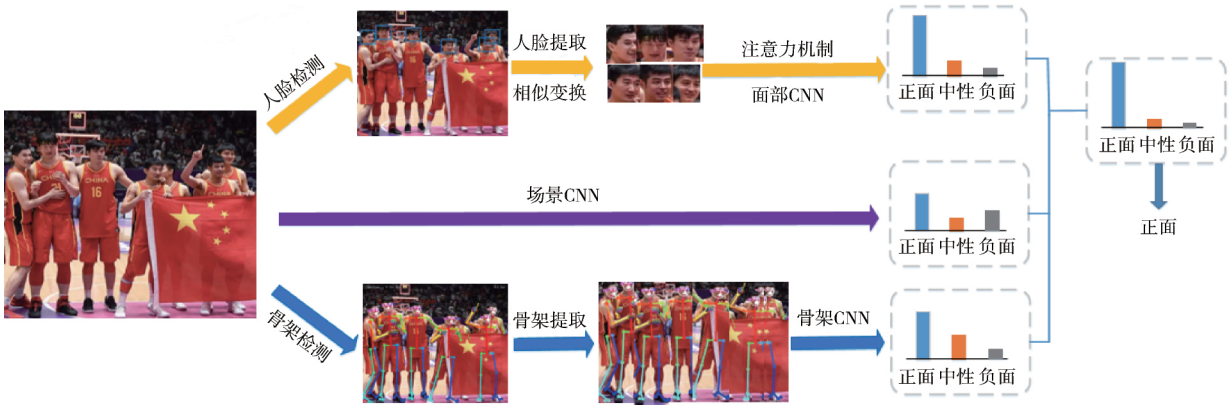

图 1 群体情绪识别系统框架

1.1.1 L-Softmax 损失函数

Softmax Loss 函数经常在卷积神经网络被用到，较为简单实用，但是它并不能够明确引导网络学习区分性较高的特征[6]。Large-margin Softmax Loss（L-Softmax）被引入用于判别学习，它能够有效地引导网络学习使得类内距离较小、类间距离较大的特征[7]，图 2 从几何角度直观地表示两种损失的差别，$W_1 = W_2$ 即指等量的二分类问题。同时，L-Softmax 不但能够调节不同的间隔（margin），而且能够减轻过拟合问题。在微调阶段，对于面部特征 x_i，损失通过以下公式计算

$$L_i = -\log \frac{e^{\|w_{y_i}\|\|x_i\|\varphi(\theta_{y_i})}}{e^{\|w_{y_i}\|\|x_i\|\varphi(\theta_{y_i})} + \sum_{j\neq y} e^{\|w_{y_i}\|\|x_i\|\cos\theta_j}} \tag{1}$$

其中，y_i 是 x_i 的标签，w_{y_i} 是全连接层中 j 类的权重

$$\cos(\theta_j) = \frac{w_j^T x_i}{\|w_j\|\|x_i\|} \tag{2}$$

$$\varphi(\theta) = (-1)^k \cos m\theta - 2k, \theta \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right] \tag{3}$$

其中，m 是预设角度边界约束，k 是整数且 $k \in [0, m-1]$。

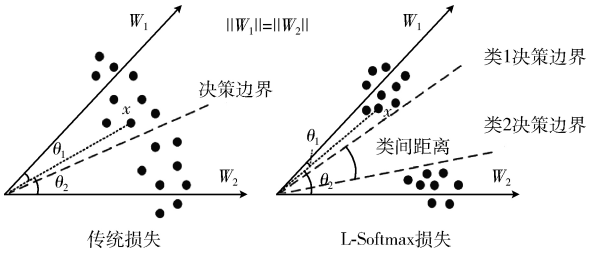

图 2 L-Softmax 损失几何理解

1.1.2 注意力机制

群体图像中存在多个人脸，为了可以独立于图像中存在的不同面部来进行情感识别，需要将所有的面部特征转换为单个表示。

最简单的解决方法是计算平均特征，但图像中某些面部情感与图像的标签无关，可能会混淆最终的分类。例如考虑哭笑的情况，许多方法容易将其混淆为负面情绪，因而无法进行有效识别。如果将置信度值与图像中每个面部相关联，就可以通过对哭泣的面部赋予较低的重要性，从而来推断图像表示正面情绪。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

基于上述理解，本文使用注意力机制来找到图像中每个面部的概率权重，根据这些权重计算加权和以产生面部特征的单个表示。该注意力机制的方案如图 3 所示。将图像中检测到的面部输入到特征提取网络，即 ResNet18。再把面部特征向量 P_i 输入到具有一维输出 μ_i 的全连接层，μ_i 获取了面部的重要性，并用其计算得分向量 P_m

$$P_m = \frac{\sum_i \mu_i p_i}{\sum_i \mu_i},\ (i = 1,2,3) \tag{4}$$

然后将 P_m 和 p_i 连接起来并将其输入另一个全连接层，其中一维输出注意权重 δ_i 表示 p_i 和 P_m 之间的关系。根据注意权重计算特征的加权和，以产生特征向量 P_d，其指示基于人脸的图像全局表示

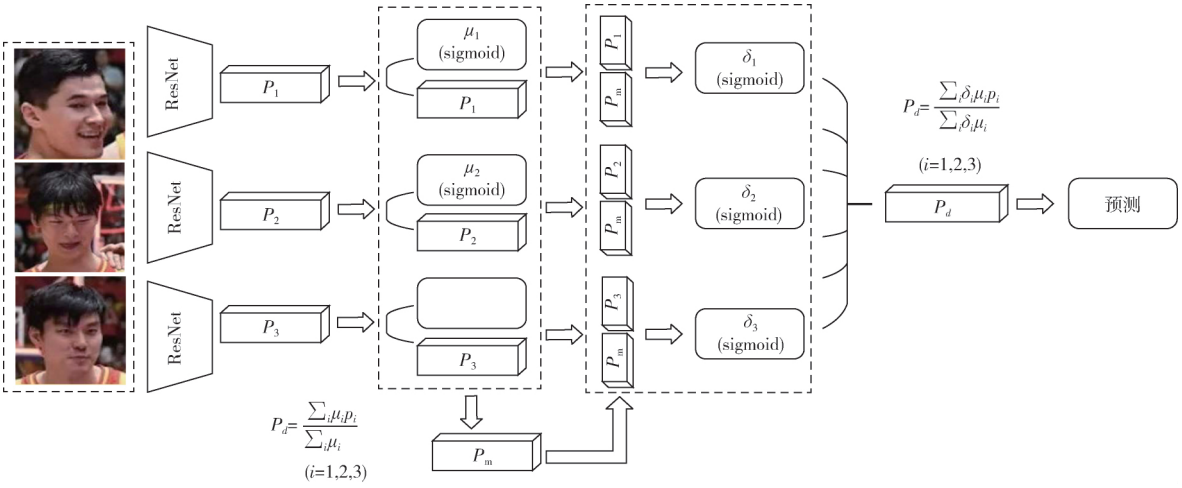

图 3　注意力机制

$$P_d = \frac{\sum_i \delta_i \mu_i p_i}{\sum_i \delta_i \mu_i},(i = 1,2,3) \tag{5}$$

1.2　场景 CNN

图像的全局场景为群体情绪识别提供重要线索。例如在葬礼期间拍摄的照片最有可能描绘出负面情绪；在婚礼中拍摄的照片最有可能表现出积极的情绪；而会议室中出现的照片更可能是中立的情绪。因此，本文使用最先进的分类网络 SE-net154[8] 从整个图像中学习全局场景特征，训练基于图像全局的场景 CNN。SE-net154 是一种先进的识别网络，引入了压缩和奖惩网络模块筛选有用特征，压缩和奖惩网络模块如图 4 所示[8]。

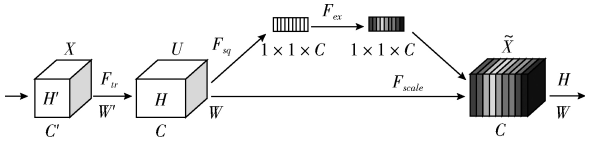

图 4　SE 网络模块

压缩和奖惩网络模块（squeeze-and-excitation blocks，SE）通过精确的建模卷积特征各个通道之间的作用关系来改善网络模型的表达能力，是一种能够让网络模型对特征进行校准的机制，使网络从全局信息出发来选择性地放大有价值的特征通道并且抑制无用的特征通道。压缩功能如下所示

$$z_c = F_{sq}(u_c) = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H} u_c(i,j) \tag{6}$$

其中，z_c 是压缩通道的第 c 个元素，$F_{sq}(.)$ 是挤压函数，u_c 是第 c 个通道的输入，W 和 H 表示输入的高度和宽度。

奖惩操作包括两个全连接层两个激活层操作，具体公式如下

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)) \tag{7}$$

其中，δ 和 σ 分别是激活函数 $ReLU$ 和 $Sigmoid$，降维层 $W_1 \in R^{\frac{c}{r}\times c}$ 和升维层 $W_2 \in R^{c\times\frac{c}{r}}$。

1.3　骨架 CNN

以往的情绪识别广泛使用人体面部特征，但根据实验心理学和情感计算的研究结果，身体姿势特征也传达重要的情感信息。为了保留人脸标志和身体特征关键点的相对位置，本文使用骨架特征表示，对应于人脸、身体和手的关键点集合。

本文使用 OpenPose[9] 来获得人体骨架姿势，它可以联合检测单幅图像中人体、手和面部的关键点（每个人总共 135 个关键点），并且与图像中检测到的人数相同，效果如图 5 所示。提取结果显示清晰的嘴形、身体姿势、手势和人物在图像中的布局，骨架特征图像与原始图像尺寸相同，再将图像按人体骨架外部最大矩形裁剪。本文使用 ResNet101[4]、SE-net154[8] 作为骨架 CNN 来识别群体情绪，首先通过模型获取图像中每个人骨架的得分，然后将所有骨架的得分和平均作为整个图像的预测。

2　实验与分析

2.1　群体图像数据库

实验使用 EmotiW 数据库，其图像来自 Group Affect

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

图 5 骨架提取

Database 2.0[5]。它包括 9815 个训练图像，4346 个验证图像和 3011 个测试图像，图像标签将群体情绪分类为正面，中性或负面。这些图像是从社交活动中收集的，例如聚会、结婚、派对、会议、葬礼、抗议等。该数据集的一些样本如图 6 所示。

图 6 EmotiW 数据库样本

2.2 网络的参数设置与训练

本文在基于 Python 的深度学习框架 PyTorch 环境下进行训练和测试实验。电脑系统环境如下：

（1）Ubuntu16.04×64；

（2）AMD Ryzen 5 1600 CPU；

（3）16 GB 内存；

（4）NVIDA GeForce GTX 1080。

2.2.1 面部 CNN 训练

本文使用多任务级联卷积网络模型（MTCNN）来检测图像中人的面部，MTCNN 是基于卷积神经网络的人脸检测方法，具有性能高和速度快的优点。它包含 3 个级联 CNN，可以快速准确地检测和对齐面部 5 个关键点（即两只眼睛、两个嘴角和鼻子）。它根据输入图像构建多尺度图像金字塔，然后将它们提供给以下三级级联框架，候选区域在第一阶段产生并在后两个阶段细化，面部标志位置在第三阶段产生。

从 MTCNN 模型获得的面部因图像差异而具有不同的方向和比例，为了学习更简单的模型，将每个面部标准化为正面视图并且统一面部图像的分辨率。可使用 5 个检测到的面部标志点来进行相似变换，使得各脸部的眼睛处于同一水平并将图像尺寸重新缩放到 96×112，获得所有基于人脸表情面部 CNN 所需要的对齐人脸。

对基于人脸表情的面部 CNN，本文训练了注意力机制模型，即 ResNet18_Attention。模型训练设置批量大小为 16，初始学习率为 0.001，且应用学习率衰减，每 9 个时期将其除以 10，持续 27 个时期。

为了比较不同先进网络架构的性能，除了 ResNet18 模型，本文还使用了 SphereFace[10]、VGG-FACE[11] 和 SE-net154。先在 FERPlus 表达数据集上预先训练这些 CNN，然后在 EmotiW 训练数据集中使用 L-Softmax 损失对它们进行微调。

2.2.2 场景 CNN 训练

对基于图像全局的场景 CNN，本文使用了 4 种网络：VGG19[11]，ResNet101[4]，SE-net154[8] 和 DenseNet-161[12]。其中 VGG19 在 Places 数据集上预先训练，ResNet101、SE-net154 和 DenseNet-161 在 ImageNet 数据集上进行预训练，然后使用 Softmax 损失在训练数据集中进行微调。在这 4 个模型中，将所有图像保持长宽比例缩放至最小边 256，这样可以最大程度保持图片形状，并随机裁剪 224×224 区域。训练参数设置与基于注意力机制的面部 CNN 模型相同。

2.2.3 骨架 CNN 训练

对于骨架 CNN，本文采用的 ResNet101 和 SE-net154 在 ImageNet 数据集上进行了预训练，然后在提取的骨架图像上进行微调，且使用与基于图像全局的场景 CNN 模型相同的训练策略。

2.2.4 模型融合

单个分类器通常不能处理现代模式识别任务的多样性和复杂性，而且决策融合不同分类器的优越性也已经得到了证明。

混合网络是通过融合各个模型的预测而构建的，在所有模型的预测中执行网格搜索以学习每个模型的权重[13]。尽管它只是通过手动指定的超参数空间子集进行穷举搜索，并且不能保证是最优的，但它是决策融合有效且广泛使用的方法。权重范围从 0 到 1，增量为 0.05，其总和限制为 1。权重为 0 的模型是冗余的，因此从混合网络中删除。

2.3 实验结果分析

评估人脸表情的面部 CNN，表 1 显示了 EmotiW 验证集上 5 种面部 CNN 模型的结果，所有型号的准确度均达到 70%左右。如表可得使用注意机制的网络比 Resnet18 基线提高了性能约 2%，即训练面部 CNN 时，本文使用注意机制是有效的。

评估基于图像全局的场景 CNN，表 2 列出了 EmotiW 验证集上 4 种场景 CNN 模型的结果。其中 VGG19 使用 L-Softmax 损失，ResNet101、SE-net154 和 DenseNet-161 使用 Softmax 损失。由表可见 SE-net154 和 DenseNet-161 获得了较优的性能。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1　EmotiW 验证集上面部 CNN 模型的结果

模型	准确率/%
SphereFace	70.94
ResNet18	69.65
ResNet18_Attention	72.12
VGG-FACE	70.89
SE-net154	71.16

表 2　EmotiW 验证集上场景 CNN 模型的结果

模型	准确率/%
VGG19	72.60
ResNet101	71.85
SE-net154	74.62
DenseNet-161	74.92

评估基于人物的骨架 CNN，表 3 显示了 EmotiW 验证集上两种骨架 CNN 模型的结果。由表可见 SE-net154 的性能优于 ResNet101。

表 3　EmotiW 验证集上骨架 CNN 模型的结果

模型	准确率/%
ResNet101	69.23
SE-net154	70.87

如图 7 给出了 3 个 CNN 分支上最优模型的混淆矩阵，可知面部 CNN 和骨架 CNN 对于正类和负类表现相对更好，但在识别中性类时更差。这背后的原因可能是这两个分支的群体情绪由人体的面部和肢体语言主导，而没有考虑人物所处的环境。场景 CNN 在识别中性时取得了较好的效果，因此有必要结合多个分支的优点，提高准确率。

混合网络最终由 7 个模型组成：SphereFace、ResNet18_Attention、ResNet18、VGG-FACE、SE-net154（场景）、DenseNet-161（场景）和 SE-net154（骨架）。表 4 显示了 EmotiW 测试集上具有不同权重的多个模型组合结果，并与文献［14］和文献［15］进行比较，准确率分别提高了 3.82%和 1.9%，验证了本文方法的有效性。

表 4　EmotiW 测试集上混合模型的结果

方法	准确率/%
本文	82.80
文献[14]	78.98
文献[15]	80.90

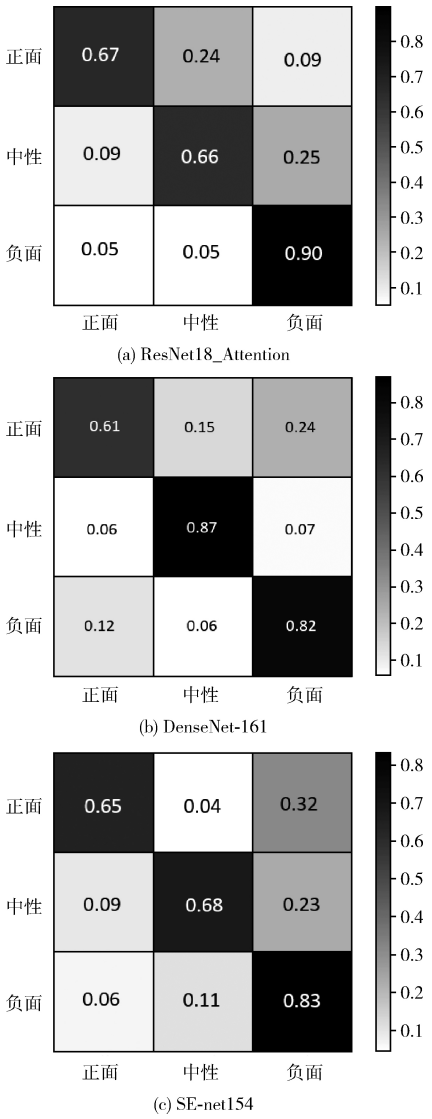

图 7　各分支最优模型的混淆矩阵

3　结束语

本文研究了对图像中小群体情绪进行识别的问题，提出基于 3 种视觉特征的卷积神经网络（CNN），即基于人脸表情的面部 CNN，基于图像全局的场景 CNN 和基于人体姿势的骨架 CNN。在面部 CNN 中引入有效的注意机制来融合不同人脸的面部特征，降低了由个别面部表情混淆最终分类的可能，还利用 L-Softmax 损失进行判别性学习。用 OpenPose 获得图像中的人体骨架作为骨架 CNN 的输入，充分利用了图像中人物的情绪线索，并在网络中引入压缩和奖惩网络模块来改善模型的表达能力。最后还探索了多个模型的决策融合，在 EmotiW 数据库上进行实验以评估所提出方法的识别性能。与现有技术方法相比，实验结果表明本文模型获得了较高的识别率，验证了本文方法的有效性。

参考文献：

［1］QING Linbo，XIONG Wenshi，ZHOU Wenjun，et al. Group

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

emotion recognition based on multi-stream CNN-LSTM network [J]. Application Research of Computers, 2018, 35 (12): 3828-3831 (in Chinese). [卿粼波, 熊文诗, 周文俊, 等. 基于多流 CNN-LSTM 网络的群体情绪识别 [J]. 计算机应用研究, 2018, 35 (12): 3828-3831.]

[2] Dhall A, Goecke R, Gedeon T. Automatic group happiness intensity analysis [J]. IEEE Transactions on Affective Computing, 2015, 6 (1): 13-26.

[3] Tan Lianzhi, Zhang Kaipeng, Wang Kai, et al. Group emotion recognition with individual facial emotion CNNs and global image based CNNs [C] //Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017: 549-552.

[4] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [EB/OL]. [2015-12-10]. https://arxiv.org/abs/1512.03385.

[5] Emad Barsoum, Zhang Cha, Cristian Canton-Ferrer, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution [EB/OL]. [2016-09-24]. https://arxiv.org/pdf/1608.01041.

[6] YU Chengbo, TIAN Tong, XIONG Di'en, et al. Face recognition under the joint supervision of central loss and Softmax loss [J]. Journal of Chongqing University, 2018, 41 (5): 92-100 (in Chinese). [余成波, 田桐, 熊递恩, 等. 中心损失与 Softmax 损失联合监督下的人脸识别 [J]. 重庆大学学报, 2018, 41 (5): 92-100.]

[7] Liu Weiyang, Wen Yandong, Yu Zhiding, et al. Large-margin softmax loss for convolutional neural networks [C] //International Conference on Machine Learning, 2016: 507-516.

[8] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks [C] //Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[9] Cao Z, Simon T, Wei S, et al. Realtime multi-person 2D pose estimation using part affinity fields [C] //Conference on Computer Vision and Pattern Recognition, 2017: 7291-7299.

[10] Liu Weiyang, Wen Yandong, Yu Zhiding, et al. SphereFace: Deep hypersphere embedding for face recognition [C] //Conference on Computer Vision and Pattern Recognition, 2017: 212-220.

[11] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2014-09-04]. https://arxiv.org/abs/1409.1556.

[12] Huang Gao, Liu Zhuang, Kilian. Densely connecte convolutional networks [C] //Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[13] WEN Bowen, DONG Wenhan, XIE Wujie, et al. Stochastic forest parameter optimization based on improved grid search algorithm [J]. Computer Engineering and Applications, 2018, 54 (10): 154-157 (in Chinese). [温博文, 董文瀚, 解武杰, 等. 基于改进网格搜索算法的随机森林参数优化 [J]. 计算机工程与应用, 2018, 54 (10): 154-157.]

[14] Guo Xin, Zhu Bin, Luisa F Polanía, et al. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions [C] //International Conference on Multimodal Interaction, 2018: 635-639.

[15] Gupta, Aarush and Agrawal, Dakshit and Chauhan, et al. An attention model for group-level emotion recognition [C] //International Conference on Multimodal Interaction, 2018: 630-634.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net