计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):15−24 [doi: 10.15888/j.cnki.csa.008707]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于局部流形注意力的人脸表情识别①

杜洋涛[1], 杨鼎康[1], 翟　鹏[1,3,5], 张立华[1,2,3,4,5]

[1](复旦大学工程与应用技术研究院, 上海 200433)
[2](季华实验室, 佛山 528200)
[3](智能机器人教育部工程研究中心, 上海 200433)
[4](吉林省人工智能与无人系统工程研究中心, 长春 130703)
[5](吉林省智能科学与工程联合重点实验室, 长春 132606)
通信作者: 张立华, E-mail: lihuazhang@fudan.edu.cn

摘　要: 人脸表情识别在各种人机交互场景中有广泛的应用, 但在表情模糊或存在遮挡情况下, 现有的表情识别方法效果并不理想. 针对表情模糊和遮挡问题, 本文提出了一种基于局部流形注意力 (SPD-Attention) 的网络架构, 利用流形学习得到具有更强描述能力的二阶统计信息以加强对表情细节特征的学习, 抑制遮挡区域无关特征对网络的影响. 同时, 针对流形学习过程中由于对数计算导致的梯度消失和爆炸, 本文提出了相应的正则约束加速网络收敛. 本文在公开表情识别数据集上测试了算法效果, 与 VGG 等经典方法相比取得了显著提升, 在 AffectNet、CK+、FER2013、FER2013plus、RAF-DB、SFEW 上正确率分别为: 57.10%、99.01%、69.51%、87.90%、86.63%、49.18%, 并在模糊、遮挡表情数据集上相比于 Covariance Pooling 等目前先进方法提升了 1.85%.

关键词: 表情识别; 流形学习; 注意力机制; 模糊遮挡表情; 卷积神经网络

引用格式: 杜洋涛,杨鼎康,翟鹏,张立华.基于局部流形注意力的人脸表情识别.计算机系统应用,2022,31(10):15–24. http://www.c-s-a.org.cn/1003-3254/8707.html

Local-manifold Attention for Facial Expression Recognition

DU Yang-Tao[1], YANG Ding-Kang[1], ZHAI Peng[1,3,5], ZHANG Li-Hua[1,2,3,4,5]

[1](Academy for Engineering & Technology, Fudan University, Shanghai 200433, China)
[2](Ji Hua Laboratory, Foshan 528200, China)
[3](Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai 200433, China)
[4](Artificial Intelligence and Unmanned Systems Engineering Research Center of Jilin Province, Changchun 130703, China)
[5](Jilin Provincial Joint Key Laboratory of Intelligent Science and Engineering, Changchun 132606, China)

Abstract: Facial expression recognition (FER) has various applications in human-computer interaction scenarios. However, existing FER methods are not that effective for blurred and occluded expression. To cope with facial expression blur and occlusion, this study proposes a novel network based on local manifold attention (SPD-Attention), which uses manifold learning to obtain the second-order statistical information with a stronger descriptive ability for strengthening the learning of facial expression details and suppressing the influence of irrelevant features in the occlusion area on the network. At the same time, in view of the disappearance and explosion of gradient caused by logarithmic calculation, this study proposes corresponding regular constraints to accelerate network convergence. The effect of the algorithm is tested on public expression recognition data sets, which is significantly improved compared with those of classic methods such as VGG. The accuracy is 57.10%, 99.01%, 69.51%, 87.90%, 86.63%, and 49.18% on AffectNet, CK+, FER2013, FER2013plus, RAF-DB, and SFEW, respectively. In addition, compared with state-of-the-art methods such as Covariance Pooling, the proposed method has an accuracy improved by 1.85% on a special blurred and occluded expression data set.

① 基金项目: 科技创新 2030-“新一代人工智能”重大项目 (2021ZD0113500)
收稿时间: 2022-01-11; 修改时间: 2022-01-30; 采用时间: 2022-02-15; csa 在线出版时间: 2022-06-24

Key words: facial expression recognition; manifold learning; attention mechanism; blurred and occluded expression; convolutional neural network (CNN)

面部表情是最自然、通用的人类情感信息传达的方式之一. 机器学习任务中, 面部表情识别可以帮助机器更好的理解人类的行为和与人类交互, 在人机协同、自动驾驶等领域有重要应用. 目前表情识别算法可以在大部分场景下发挥较好效果, 但是对于模糊微弱表情或伴随着遮挡、面部姿势、光照等外在干扰情况, 现有算法仍然具有一定缺陷.

一般来说, 面部表情识别主要包括 3 个阶段, 即人脸检测、特征提取和表情分类. 在人脸检测阶段, Dlib[1] 是较为轻量级别的检测接口, 但由于其对侧脸检测效果一般, 所以 MTCNN[2] 人脸检测器在复杂场景中更为常用. 表情的特征提取工作可以根据特征类型, 分为手工设计特征和基于学习得到的特征. 手工特征主要分为基于纹理表达和基于几何结构两类, 如 SIFT[3]、HOG[4]、LBP 直方图[5]、Gabor 小波系数[6] 等属于经典的纹理表达特征, 同时还有大量研究基于鼻子、眼睛和嘴巴周围关键点的相关几何特征. 随着 GPU 等并行计算设备高速发展, 深度学习逐渐成为图像分析领域的主要研究方向, 因此基于学习得到特征是目前表情特征提取的主流方法. Tang[7] 利用卷积神经网络 (convolutional neural networks, CNN) 进行特征提取并分类, Liu 等人[8] 也提出了一种基于面部动作单元的 CNN 架构用于表情识别.

大部分现实生活中的表情并不会像实验室采集的数据那样具有明显的表情幅度特征, 更多以微弱表情形式呈现, 同时真实场景下人脸表情关键部位可能会被墨镜、口罩、帽子等遮挡. 模糊微弱表情识别的主要难点在于表情特征过小会被淹没与人脸图像特征中, 使得表情特征无法对分类器训练产生促进效果. 为了解决此问题, Peng 等人[9] 基于迁移学习的概念, 在经过 ImageNet 预训练的 ResNet101 网络基础上, 用表情数据集进行冻结训练微调参数. Khor 等人[10] 进一步提出丰富的长期递归卷积网络 (enriched long-term recurrent convolutional network, ELRCN), 在长短期记忆网络 (long short-term memory, LSTM) 结构基础上利用 CNN 模块完成模糊表情序列的特征向量编码, 最后实现分类. Peng 等人[11] 和 Huang 等人[12] 利用表情序列进行特征增强, 通过计算表情幅度最大的图像帧与序列中某一代表帧的光流特征作为时序信息, 然后利用 CNN 完成模糊表情识别, 同时提高了分类正确率和节省了序列信息的计算耗时.

面部遮挡、光照明暗和不同的姿势带来的面部表情信息损失通常发生在现实世界的场景, 因为面部区域可以很容易被太阳眼镜, 帽子, 围巾等遮挡. Liu 等人[13] 提出利用 Gabor 直方图衡量图像部分遮挡, 并引入 LGBPHS 方法来解决. Cotter[14,15] 提出对部分遮挡图片使用稀疏性表示分类器效果较差的问题. Li 等人[16] 设计了一个基于补丁的注意网络, 用于遮挡感知下的表情识别. 对于位姿变化问题, Rudovic 等人[17] 提出了耦合比例高斯过程回归 (CSGPR) 模型的头部归一化, Lai 等人[18] 利用 GAN 从侧脸图像生成正脸图像来解决面部姿势问题.

由于表情模糊导致的表情特征不明显和由于遮挡导致的表情特征不可见是表情识别领域的两个重要问题. 因此如果直接对面部图像进行特征提取, 微弱模糊的表情特征容易被忽略, 同时遮挡区域则会提取面部无关特征, 而人类却可以“放大”微弱表情特征同时忽略遮挡区域的无关特征. 心理学研究表明[19], 人类的注意力机制可以有效地利用局部区域和整体面孔来感知不完整面孔传递地语义信息. 受到此研究启发, 近年来涌现出许多基于注意力机制 (attention) 的深度学习方法.

注意机制是在强化算法的基础上发展起来的, 但却广泛地应用于视觉深度学习领域的局部特征强化. Badanau 等人[20] 同时对翻译和源语言对齐两项任务进行注意力运算, 他们的工作是首次尝试将注意机制应用于机器翻译, 并获得突破性结果. 随后, 注意力模型在深度学习领域被广泛应用, 针对不同任务提出了多种注意机制模型, 如针对机器阅读的 LSTM模型、机器翻译[21] 的多类注意模型和视频分类[22] 的注意集群模型. 在计算机视觉领域, 注意力模型也取得了异常成功的效果, Wang 等人[23] 提出了一种人脸检测的注意网络, 在生成锚点的步骤中突出了人脸区域. Yang 等人提出的神经聚合网络 (neural aggregation network, NAN)[24], NAN 使用级联注意机制将一个视频或集合

的人脸特征聚合成一个紧凑的视频表示.

上述讨论的所有深度学习网络几乎均采用传统的卷积、池化、全连接等网络层. Yu 等人研究认为[25], 传统的卷积神经网络 (CNNs) 使用卷积层、最大池化或平均池化和全连接层只能捕获一阶统计量, 而二阶统计量如协方差等被认为是比第一阶统计量 (如均值或最大值) 更好的图像区域语义描述符[26]. 而基于流行网络的特征提取模块可以捕获二阶统计量, 更好的刻画图像扭曲的特征. 在文献 [25,27,28] 中, 作者基于 VGG 网络的各种架构实验二阶特征的合并效果, 并在图像分类、目标检测等数据集上进行实验. 而在表情识别领域, Acharya 等人在文献 [28] 中提出了一种协方差池化 (covariance pooling) 的深度学习架构, 该工作分析了驻留在 SPD 流形[29,30] 上的二阶统计特征, 并构造成网络框架对特征协方差进行学习迭代, 其实验表明在表情识别任务中拥有较好的效果. 但是协方差池化往往针对全局采用相同的计算系数, 对于表情微弱和局部遮挡等情况并未做考虑, 因此本文改进其方法, 以更好的应对模糊微弱表情和遮挡情况.

因此本文认为二阶统计特征可以更好地描述区域的扭曲程度从而更好地学习表情语义, 如果结合注意力机制将更有效地提取局部微弱表情特征同时抑制无关特征. 因此本文利用流形学习网络获取局部区域的二阶统计特征并将其作为局部注意力特征输入主网络中. 值得指出的是, 本文并非第一个提出用局部注意力机制解决表情识别问题的, 但尽我们所知, 本文是第一个用流形学习获取二阶统计信息作为面部区域注意力系数的.

综上所述, 本文的主要贡献有:

(1) 构建了面部表情局部注意力网络框架, 利用注意力机制强化模糊微弱表情等情况下的微小表情特征, 同时抑制由于墨镜、口罩等面部遮挡物带来的表情无关特征, 从而提高表情识别能力;

(2) 提出流形注意力机制模块 (SPD-Attention module), 构造对称正定的协方差矩阵结合流形学习网络得到二阶统计特征刻画局部区域的扭曲程度, 相比于一阶特征可以更好地刻画表情特, 同时提出了对流形网络过程的正则化损失, 提高其收敛速度;

(3) 在 AffectNet、CK+、FER2013、FER2013plus、RAF-DB、SFEW 和公开的模糊遮挡数据集上进行了测评, 相比与 ResNet34、VGG19 等经典深度学习方法具有普遍提升效果, 同时与目前先进方法对比也取得了近似或更好的水平.

1 基于局部流形注意力的表情识别

目前大部分表情识别方法主要基于深度学习网络, 对于微表情和遮挡问题也主要依赖于网络结构的调整和细化. 基于以上研究方向, 本文希望可以通过注意力机制自动地增强较小的表情特征和抑制遮挡带来的无关特征, 并构建一个端到端 (end-to-end) 的网络结构进行学习训练.

由于注意力机制可以由不同的网络结构实现, 其本质是对特征的赋权, 如文献 [31] 中, Wang 等人利用共享的全连接层训练注意力系数. 但考虑到二阶统计信息往往可以更好地刻画面部扭曲情况 (而表情语义往往蕴含与局部扭曲), 因此本文考虑采用二阶统计信息来构造注意力机制, 设计算法 1.

算法 1. 基于局部流形注意力的表情识别框架

1) 从将输入图片$image_{\text{raw}}$进行一份拷贝和 n 份局部剪裁构成图片集$S=\{image_0,image_1,\cdots,image_n\}$;

2) 将图片集S输入一个共享权值的基础卷积神经网络 (CNN), 并提取该网络某一层 (本文提取倒数第 2 层) 的特征图集合$O=\{output_0,output_1,\cdots,output_n\}$和该网络最后的特征向量集合$F=\{fea_0,fea_1,\cdots,fea_n\}$;

3) 将每个特征图$output_i$和特征向量fea_i输入流形注意力计算模块 (SPD-Attention module) 中得到带注意力的 j 局部特征向量$vector_i$, 最后求和输出$\widetilde{Vector}$;

4) 计将求和后的特征向量输入全连接层FC_1中得到分类预测结果.

这样做的好处主要有两点: (1) 模糊微弱表情的二阶统计特征相对一阶统计特征更明显, 可以提高分类效果; (2) 面部图像中的墨镜等遮挡物在二阶统计信息会被相对抑制, 可以过滤表情无关特征. 本文的网络结构如图 1 所示.

1.1 局部注意力机制网络

如图 1 所示, 本文的主体框架是基于局部图像的自注意力机制网络结构, 主要包含 3 个步骤, 第 1 步是获得表情图像的联合特征向量: 输入原始图像的一份拷贝和 n 份局部剪裁图片构成整体输入图片集合$S=\{image_0,image_1,\cdots,image_n\}$, 经过基础特征提取卷积神经网络 (该网络对$S$中的每个元素权值共享) 得到特征图和特征向量; 第 2 步是利用第 1 步得到的特征图求取协方差矩阵, 然后经过流形注意力模块 (SPD-Attention module) 得到带注意力的图像待分类特征向量; 第 3 步则是利用最终的待分类特征向量通过全连接层进行分类预测. 我们将在第 1.2 节详细介绍 SPD-Attention 模块, 此处将详细介绍第 1 步和第 3 步.

为了保证网络可以得到面部表情的全局特征作为

局部特征的参考, 对于输入图片$image_{raw}$, 我们首先复制其本身得到$image_0$, 然后对其进行局部剪裁得到局部图像序列$image_1, image_2, \cdots, image_n$本文采用的局部剪裁方法主要有两种, 第 1 种是随机剪裁: 设定剪裁区域面积占总面积的比例为 r (本文取 r=0.75), 然后随机选取相应面积区域; 第 2 种是面部关键点剪裁: 根据标定的 83 个人脸关键点[2], 剪裁关键点周围区域. 剪裁完成后将得到的图片集$S=\{image_0, image_1, \cdots, image_n\}$输入一个权值共享的 CNN 网络中, 本文采用 ResNet18 作为基本框架, 得到特征图输出集合$O=\{output_0, output_1, \cdots, output_n\}$和最后的特征向量集合$F=\{fea_0, fea_1, \cdots, fea_n\}$.

上述特征图和特征向量经过局部流行注意力机制后再求和得到最终的待分类特征向量$\widetilde{Vector}$并将其输入最后一层全连接层FC_1得到分类的预测结果, 由于是多分类任务, 本文采用的分类损失函数为交叉熵损失 (CrossEntropy), 具体如式 (1) 所示:

$$L_{CE}=-\sum_{k=1}^{N}(p_k \log q_k) \tag{1}$$

其中, p 为标签值, q 为预测值 (经过 Softmax 后的 one-hot 形式).

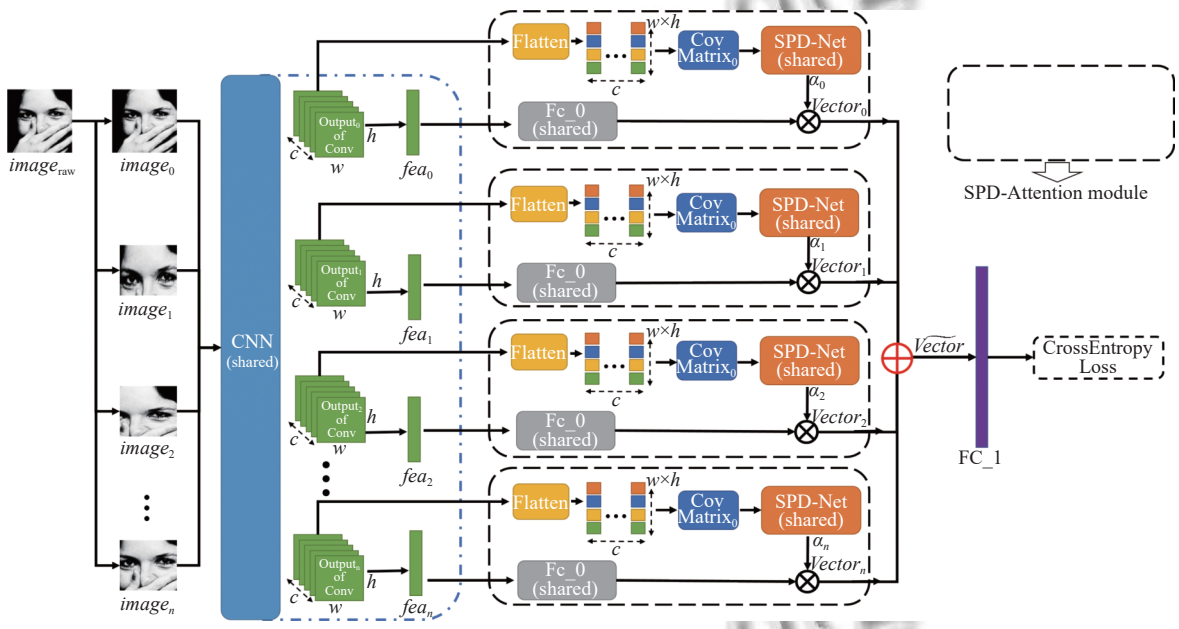


图 1 基于局部流形注意力的表情识别网络框架

1.2 SPD 流形注意力机制

流形注意力机制 (SPD-Attention module) 是本文的核心创新点, 如图 1 虚线框中模块所示. 对于 CNN 输出的特征图输出集合$O=\{output_0, output_1, \cdots, output_n\}$中的每个元素$output_i$进行相同的操作, 首先将$output_i$的每一层拉平, 假设原来的$output_i$的维度为$w\times h\times c$, 拉平后维度变成$wh\times c$即变成$c$个向量组, 由此可以对它们求协方差矩阵$Covmatrix_i$, Tuzel 等人[26] 的研究表明由特征计算出的协方差矩阵驻留在 SPD 流形上, 其相较于一阶统计信息可以更好地捕获区域特征. 值得一提的是, CNN 输出的特征图理论上可以选取任一层, 本文选取 ResNet18 的倒数第 2 层作为输出.

假设特征图拉平后得到的向量为$\{x_1, x_2, \cdots, x_c\}\in\mathbb{R}^{wh}$则协方差矩阵为式 (2):

$$Covmatrix_i=\frac{1}{c-1}(x_j-\bar{x})(x_j-\bar{x})^{\mathrm{T}} \tag{2}$$

其中, $\bar{x}$为均值.

当向量集$\{x_1, x_2, \cdots, x_c\}$中的线性独立元素个数大于$wh$时, 该协方差矩阵为对称正定矩阵 (SPD), 而只有在协方差矩阵满足 SPD 性质时, 黎曼流形的 SPD 网络结构[29] 才得以使用. 而协方差矩阵一定满足对称性, 根据式 (3) 可以证明协方差矩阵$Covmatrix_i$一定半正定.

$$\begin{aligned}\beta^{\mathrm{T}}Cov\beta&=\beta^{\mathrm{T}}E[(X-\mu)(X-\mu)^{\mathrm{T}}]\beta\\&=E[\beta^{\mathrm{T}}(X-\mu)(X-\mu)^{\mathrm{T}}\beta]=E[s^2]\geqslant 0\end{aligned} \tag{3}$$

其中, β为任意向量.

因此可以通过矩阵迹的方式将协方差矩阵正定化,

即如式 (4) 所示:

$$Covmatrix_i^+ = Covmatrix + \lambda trace(Covmatrix)I \quad (4)$$

其中, λ为正则系数, I为单位矩阵.

Huang 等人[30] 提出了 Bilinear Mapping 层、Eigenvalue Rectification 层、Log Eigenvalue 层可以在黎曼流形空间进行参数学习, 本文在其基础上对驻留在 SPD 流形上的协方差特征矩阵进行了网络训练, 下文将简要介绍, 流形学习流程如图 2 所示.

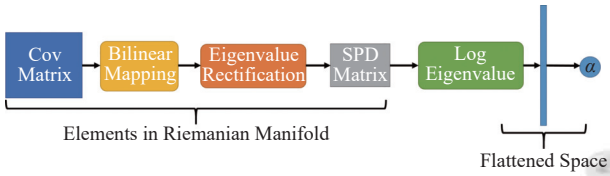


图 2 SPD 流形学习流程

由于直接将 CNN 输出的特征图拉平计算协方差矩阵 Cov, 因此 Cov 维度可能非常大, 而且可能并不适合用传统的网络层连接方式. 所以采用 Bilinear Mapping 层代替传统的网络层链接方式, 可以在降低维度的同时保证其几何结构不变, 其具体如式 (5) 所示:

$$X_k = f_b^k(X_{k-1}, W_k) = W_k X_{k-1} W_k^{\mathrm{T}} \quad (5)$$

其中, X_{k-1}为输入 Bilinear Mapping 层的 SPD 矩阵, W_k为权值矩阵, X_k为输出的 SPD 矩阵.

传统的 CNN 在每一层卷积池化之后往往需要添加如 ReLU 等激活函数层, 而在黎曼流形下可以采用 Eigenvalue Rectification 层代替, 其具体如式 (6) 所示:

$$X_k = f_r^k(X_{k-1}) = U_{k-1} \max(\varepsilon I, \sigma_{k-1}) U_{k-1}^{\mathrm{T}} \quad (6)$$

其中, X_{k-1}为输入 Eigenvalue Rectification 层的 SPD 矩阵, X_k为输出的 SPD 矩阵, ε为阈值, U_{k-1}和Σ_{k-1}为X_{k-1}的矩阵特征向量和特征值, 即$X_{k-1} = U_{k-1}\Sigma_{k-1}U_{k-1}^{\mathrm{T}}$

由于黎曼流形于传统欧氏空间计算法则并不一致, 所以可以采用 Log Eigenvalue 层使黎曼流形中的元素具有李群结构, 其输出矩阵可以展平并且可以应用标准的欧几里得运算. 其具体如式 (7) 所示:

$$X_k = f_l^k(X_{k-1}) = \log(X_{k-1}) = U_{k-1} \log(\Sigma_{k-1}) U_{k-1}^{\mathrm{T}} \quad (7)$$

其中, X_{k-1}为输入 Log Eigenvalue 层的 SPD 矩阵, X_k为输出的 SPD 矩阵, U_{k-1}和Σ_{k-1}为X_{k-1}的矩阵特征向量和特征值, 即$X_{k-1} = U_{k-1}\Sigma_{k-1}U_{k-1}^{\mathrm{T}}$.

1.3 正则损失

如第 1.2 节中式 (7) 所示, 由于在网络末端的全链接结构需要进行标准的欧几里得运算, 所以需要采用 Log Eigenvalue 层将分布于黎曼空间的 SPD 矩阵转到欧氏空间中. 虽然 Eigenvalue Rectification 层保证了Σ_k中的数值符合 Log Eigenvalue 层的计算定义, 但是由于对数运算本身的性质, 当$\Sigma_k(i,i)$趋向于 0 时容易导致梯度爆炸, 而$\Sigma_k(i,i)$过大时容易导致梯度消失. 为了网络收敛的稳健性, 本文此处引入正则项, 约束Σ_k的分布, 保证良好的梯度性质.

根据泰勒一阶展开, 我们可以定义正则约束如式 (8) 所示:

$$L_{\log\mathrm{reg}} = \frac{1}{c}\sum (\Sigma_k(i,i) - 1 - \log(\Sigma_k(i,i)))^2 \quad (8)$$

根据 Log Eigenvalue 层的 Backpropagation 规则[30], 我们可以修正梯度式 (9) 到式 (10).

$$\frac{\partial L^{(k)}}{\partial \Sigma} = \Sigma^{-1} U^{\mathrm{T}} \frac{\partial L^{(k+1)}}{\partial X_k} U \quad (9)$$

$$\frac{\partial L^{(k)}}{\partial \Sigma} = \Sigma^{-1} U^{\mathrm{T}} \frac{\partial L^{(k+1)}}{\partial X_k} U + (I - \Sigma^{-1}) \quad (10)$$

其中, $L^{(k)}$为第 k 层的 loss, I为单位矩阵.

2 实验与结果分析

为了充分验证算法的有效性, 本文首先在多个通用的表情识别数据集上实验了效果, 并与 VGG、ResNet 等经典深度学习方法和 Covariance Pooling[28] 等时下先进方法进行比较; 其次为了进一步验证本文算法对于微弱表情和遮挡情况的效果提升, 在专用的模糊遮挡表情数据集上验证了效果, 并与时下先进方法进行对比; 最后, 本文对具体效果进行了可视化, 对比分析了正则项对于流形网络的梯度约束作用.

2.1 实验数据与细节介绍

本文在 6 个公开通用表情数据集和 1 个公开专用遮挡或模糊表情数据集上实验了本文的效果; 6 个通用表情数据集分别是: Affectnet、CK+、FER2013、FER2013plus、RAFDB 和 SFEW, 专用遮挡或模糊表情数据集来自 Kai 等人的工作[31].

AffectNet 数据集. AffectNet 是一个由互联网图片组成了规模巨大的数据集, 其标注包含了离散的表情分和连续的 VA 标注信息. AffectNet 由一个不平衡的训练集和一个平衡的测试集组成, 在本文实验中, 采用的是其中给出 8 种基本表情类别 (分别为: 愤怒、蔑视、嫌弃、恐惧、高兴、中性、悲伤、惊讶) 的

450k 张数据作为训练集, 4k 张数据作为测试集.

CK+数据集. CK+数据集是经典的实验室数据集, 由序列图片构成, 图片序列展示了表情幅度由弱变强的过程, 其中包含了 8 种基本表情. 与目前大多数方法类似, 本文在处理 CK+数据时取最后 3 帧表情图片用于实验, 同时利用 MTCNN[2] 等算法对图片进行人脸裁剪, 除去背景等无关信息.

FER2013 数据集. FER2013 数据集是 ICML2013 的比赛数据集, 是一个大规模的现实生活环境表情数据集. FER2013 包含 28 709 张训练集, 3 589 张验证集和 3 589 张测试集, 每张图片是 48×48 的灰度图片, 共有 7 种表情标签.

FER2013plus 数据集. FER2013plus 数据集是由数据集 FER2013 扩展而成, 包含了 10 种离散的表情标签. 同 Covariance Pooling 工作一致, 我们选取其中标签位 8 种基本表情的图片为实验数据. 值得注意的是, FER2013plus 数据集并未给出唯一真实标签, 而是公布了所有标注者对同一张图片的标注信息, 本文同以往工作一致, 根据最大投票原则确定表情类别.

RAFDB 数据集. RAFDB 包含 30 000 张由受过训练的 40 位标注人员给出多重表情标签的图片, 图像质量和标签质量均相对较高, 在本文的实验中, 与 Covariance Pooling 工作一致, 采用基础的 12 271 张表情标注图片用于训练, 3 068 张作为测试.

SFEW 数据集. SFEW 数据集是数据集 AFEW 的子集, 包含 958 张训练集图片, 436 张验证集图片和 372 张测试集图片. 由于测试集标签并未开源, 所以同以往工作一致, 本文利用训练集训练, 验证集测试效果.

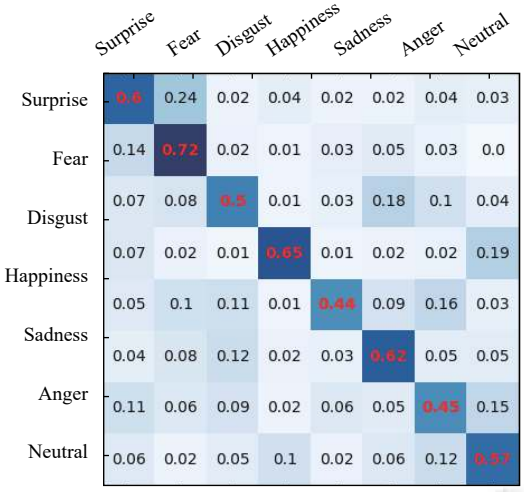
由于 SFEW 数据集过小, 为了提高训练效果, 本文将 RAFDB 数据集的训练集加入 SFEW 数据的训练过程中, 提升模型泛化能力. 由于 AffectNet 等数据集存在数据分布不平衡的问题, 同时 CK+数据集存在训练样本分布较小的问题, 所以本文对于 AffectNet 等数据集采用带权重的分类损失, 权重正比例于训练集的样本分布, 同时对 CK+等数据采用水平翻转、随机剪裁等数据增强方法扩大训练数据集规模, 图片尺寸统一采用 224×224. 本文在 PyTorch 框架下进行实验, 网络 backbone 为 PyTorch 官方提供的 ResNet18, 训练时 batchsize 为 256, 测试时 batchsize 为 128, 学习率初始为 0.1, 并每 20 个 epoch 下降 10%, 实验在 Tesla V100 GOSUs 平台完成.

2.2 通用数据集实验结果

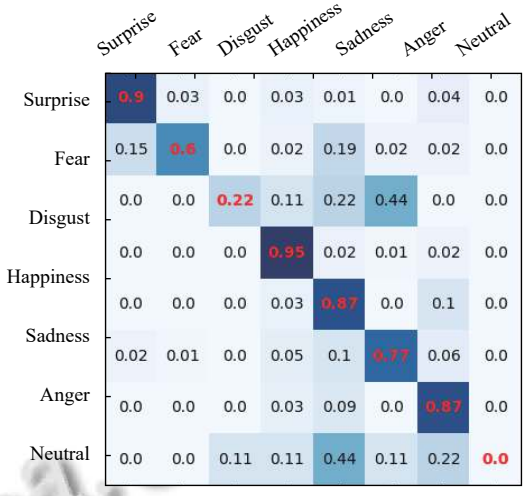
为了验证本文算法的有效性, 本文首先在 Affectnet、CK+、FER2013、FER2013plus、RAFDB 和 SFEW 6 个通用数据集上实验了本文算法, 并于经典的深度学习方法 ResNet18、ResNet34、VGG16、VGG19 进行了对比实验. 如图 3 所示, 分别展示了本文方法在 6 个通用数据集上的混淆矩阵. 从图中可以看出在本文方法在 CK+、FER2013plus 和 RAFDB 上表现较好, 而在 SFEW 上表现则一般. 这主要是因为 CK+数据集较为简单, FER2013plus 和 RAFDB 的数据较为清晰, 相比之下 SFEW 的数据难度较大, 且图像存在较多干扰, 值得指出的是, 目前的所有先进方法在 SFEW 数据上的表现均远逊于其他数据集. 同时对于不同种类的表情的识别结果也有显著区别, 对于高兴这类表情的识别效果普遍较好, 对于恐惧等表情识别效果一般较差. 这主要是因为高兴的表情具有区分性非常强的特征, 而恐惧的表情特征则容易和惊讶等表情混淆.

如表 1 所示, 本文在 6 个数据集上和经典的深度学习方法进行了对比. VGG 网络是图像识别领域非常经典的网络结构, 在人脸识别领域具有广泛应用, ResNet 由于其残差学习的特质, 可以适应于大规模深度网络, 同时也是 ImageNet 比赛的冠军网络框架. 为了更好的消融对比, 本文此处采用的 ResNet 和 VGG 均保持原有框架, 分类器统一选择全连接层, 除此之外没有其他任何模块. 为了对比的公平性, 本文此处采用相同数据集进行训练, 未加载任何人脸识别预训练权重, 优化方法均选择随机梯度下降方式 (SGD). 从表格中可以看出, 本文的方法相比与经典方法在所有数据集上均有效果提升.

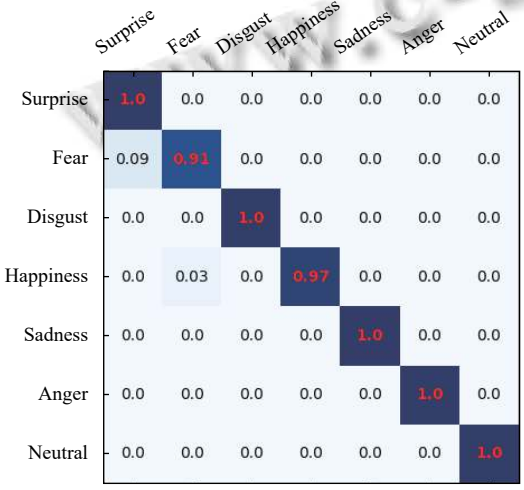
在验证效果普遍优于经典方法后, 由于目前先进方法普遍是基于 ResNet 等经典方法改进的, 所以为了进一步验证本文算法效果, 在 RAFDB 数据集上与目前先进方法进行了效果对比, 结果如表 2 所示, 此处选择 RAFDB 数据集的原因主要是由于 RAFDB 是公认的高质量数据集, 所有目前先进方法均汇报了在其上的效果. 如表 2 所示, 本文相比目前先进方法在 RADFB 数据上的效果也均有提升. 值得一提的是, 对比方法中的 Covariance Pooling 同样也采用了流形学习的方式, 利用全局的协方差池化提取二阶统计信息, 但本文方法相比于其增加了注意力机制的使用, 将二阶统计信息用作注意力系数, 更好地把握了表情的局部信息, 由此相比与 Covariance Pooling 本文有更好的识别效果.



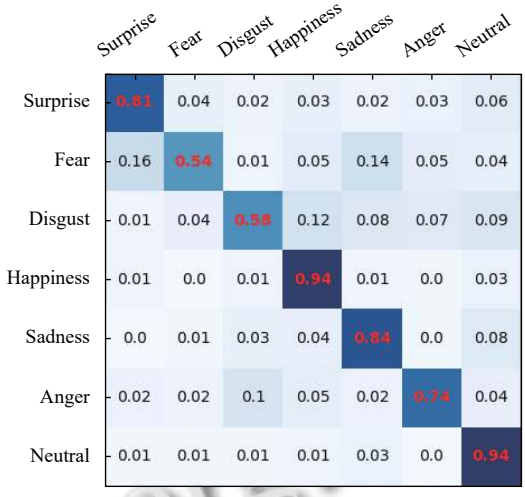
(a) SPD-Attention on AffectNet



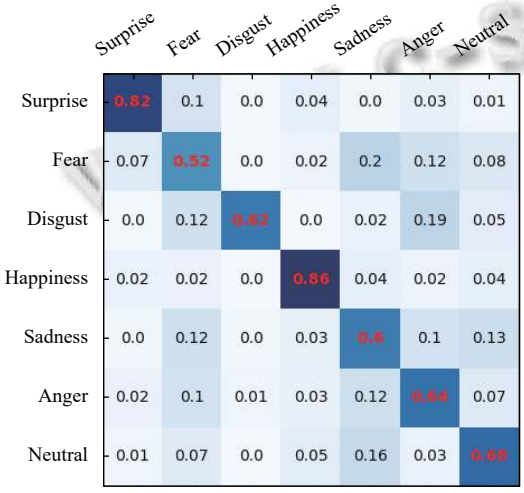
(b) SPD-Attention on CK+



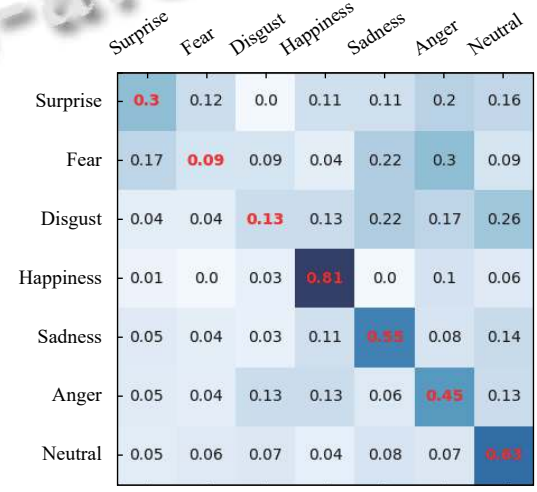
(c) SPD-Attention on FER2013



(d) SPD-Attention on FER2013plus



(e) SPD-Attention on RAFDB



(f) SPD-Attention on SFEW

图 3 通用表情识别数据集结果

表 1 通用表情识别数据集上与经典方法对比结果 (%)

方法	AffectNet	CK+	FER2013	FER2013plus	RAFDB	SFEW
ResNet18	47.51	97.40	60.30	83.22	81.21	42.10
ResNet34	52.31	98.62	60.01	85.50	80.48	38.65
VGG16	49.32	94.53	65.31	87.11	82.50	39.43
VGG19	53.06	97.32	65.23	85.95	81.01	41.20
本文方法	**57.10**	**99.01**	**69.51**	**87.90**	**86.63**	**49.18**

表 2 RAFDB 上与目前先进方法对比结果

方法	Network (backbone)	Performance (%)
Inception-ResNetV1[31]	ResNet	82.6
DLP-CNN[32]	baseDCNN	84.13
gACNN[31]	VGG	85.07
Covariance Pooling[28]	ResNet	85
本文方法	ResNet	**86.63**

注: 对比方法数据均为其论文中报告结果.

2.3 模糊、遮挡专用数据集实验结果

本文方法的设计初衷是为了应对表情数据中出现模糊微弱表情或者面部存在遮挡的情况, 所以此处本文在专用的 RAFDB 模糊、遮挡数据集[31] 上进行实验, 并于目前先进方法进行对比, 结果如表 3 所示, 混淆矩阵如图 4、图 5 所示.

表 3 模糊、遮挡表情数据集上与目前先进方法对比结果 (%)

方法	Occlusion	Pose blur
Finetune[33]	80.19	**83.15**
Covariance Pooling[28]	78.03	80.01
本文方法	**82.04**	**83.15**

注: Finetune结果数据为其论文中报告结果, Covariance Pooling结果为本文自行训练得到的结果.

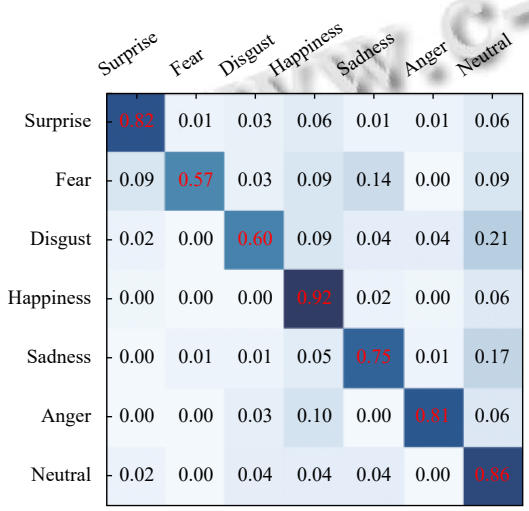


图 4 SPD-Attention on Occlusion

可以看出, 在模糊微弱表情或存在遮挡情况下本文算法相较于目前先进方法有一定提升. Covariance Pooling 方法在模糊、遮挡数据集上表现效果显著降低, 本文相较于其具有 3% 以上的提升, 主要原因是遮挡或模糊数据集主要信息来自于面部的局部信息, 而 Covariance Pooling 采用的是全局协方差池化, 缺乏对局部信息的关注. 值得一提的是, Finetune 的效果是利用专用的网络并在大规模专用数据集上进行细节化调整得到的结果, 其可以反映目前先进方法在遮挡等情况下的表情识别效果.

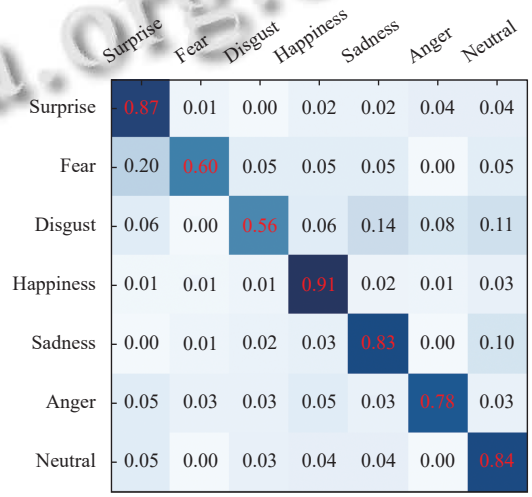


图 5 SPD-Attention on Pose blur

通过和目前先进方法的对比, 可看出本文利用流形学习得到二阶统计信息会比其他方法使用的一阶统计信息具有更好的表情特征描述能力, 也具备更好的抗干扰能力; 同时也可以看出, 相较目前利用二阶统计信息的先进工作, 本文将二阶统计信息作为注意力系数放入网络架构中, 可以更好地关注局部信息, 提高对微弱表情特征的学习, 抑制无关遮挡信息的影响.

为了进一步探究不同程度表情模糊数据对本文算法的影响, 针对 FER2013plus 数据集上不同程度姿态模糊数据[31] 进行对比实验. 如表 4 所示, 对于大姿态 (Pose45) 变化导致的表情模糊, 本文算法效果会有一定下降, 但与目前先进方法效果持平, 本文认为是大姿态模糊情况下表情信息非常有限导致.

表 4 不同程度模糊数据本文方法对比结果 (%)

方法	Pose30 blur	Pose45 blur
Finetune[33]	78.11	75.50
本文方法	**79.59**	**75.71**

注: Finetune结果数据为其论文中报告结果.

2.4 可视化实验结果

如图 6 所示, 是本文利用梯度热力图方式对本文算法的注意力区域进行可视化, 利用网络在梯度传递时对于不同区域的梯度大小, 表示网络对图像区域的关注程度. 颜色越鲜艳的区域表示网络对于该区域的梯度更大, 即表示该区域的内容对于网络识别越重要.

图 6 网络热力图可视化

从图 6 中可以看出, 本文算法着重关注面部的扭曲区域, 对于手、眼睛等遮挡具有较强的对抗能力, 同时对于模糊微弱表情也可以着重关注表情特征显著区域.

2.5 流形损失正则化效果

由于流形网络计算需要通过对数运算将 SPD 流形结构转到欧氏空间, 为了抑制梯度爆炸和梯度消失, 本文对网络的对数层引入正则损失. 由图 7 所示不同程度正则损失对于网络梯度的影响, 可以得出结论: 合适的正则损失可以有效提高网络的鲁棒性.

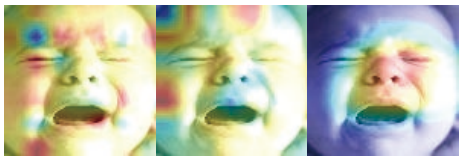


图 7 不同程度正则损失对于网络梯度的影响

3 结论与展望

本文重点关注了表情识别中模糊微弱表情和存在遮挡的表情, 从注意力机制和流形学习角度出发, 利用表达能力更强的二阶统计信息表达局部表情特征, 提出了局部流形注意力机制. 通过和 ResNet 等经典方法以及 Covariance Pooling 等目前先进方法的实验对比, 验证了本文算法对于模糊、遮挡表情识别有较好的效果. 同时为了抑制流形学习可能带来的梯度消失、梯度爆炸等情况, 利用正则损失约束网络梯度, 提高了网络稳定性. 在未来工作中, 本文将继续探究流形注意力机制在人脸识别等任务中的应用可能, 同时从知识蒸馏等角度降低由于流形网络计算带来的整体算法复杂度的提升, 探究在移动设备端使用的价值.

参考文献

1 Amos B, Ludwiczuk B, Satyanarayanan M. Openface: A general-purpose face recognition library with mobile applications. Pittsburgh: Carnegie Mellon University, 2016.

2 Zhang KP, Zhang ZP, Li ZF, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016, 23(10): 1499–1503. [doi: 10.1109/LSP.2016.2603342]

3 Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 2003, 31(13): 3812–3814. [doi: 10.1093/nar/gkg509]

4 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893.

5 Shan CF, Gong SG, McOwan PW. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 2009, 27(6): 803–816. [doi: 10.1016/j.imavis.2008.08.005]

6 Liu CJ, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing, 2002, 11(4): 467–476. [doi: 10.1109/TIP.2002.999679]

7 Tang YC. Deep learning using linear support vector machines. arXiv: 1306.0239, 2013.

8 Liu MY, Li SX, Shan SG, *et al*. Au-inspired deep networks for facial expression feature learning. Neurocomputing, 2015, 159: 126–136. [doi: 10.1016/j.neucom.2015.02.011]

9 Peng M, Wu Z, Zhang ZH, *et al*. From macro to micro expression recognition: Deep learning on small datasets using transfer learning. Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an: IEEE, 2018. 657–661.

10 Khor HQ, See J, Phan RCW, *et al*. Enriched long-term recurrent convolutional network for facial micro-expression recognition. Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an: IEEE, 2018. 667–674.

11 Peng M, Wang CY, Bi T, *et al*. A novel apex-time network for cross-dataset micro-expression recognition. Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). Cambridge: IEEE, 2019. 1–6.

12 Huang J, Zhao XR, Zheng LM. SHCFNet on micro-expression recognition system. Proceedings of the 2020 13th International Congress on Image and Signal Processing,

BioMedical Engineering and Informatics (CISP-BMEI). Chengdu: IEEE, 2020. 163–168.
13 Liu SS, Zhang Y, Liu KP, *et al*. Facial expression recognition under partial occlusion based on Gabor multi-orientation features fusion and local Gabor binary pattern histogram sequence. Proceedings of the 9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Beijing: IEEE, 2013. 218–222.
14 Cotter SF. Sparse representation for accurate classification of corrupted and occluded facial expressions. Proceedings of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas: IEEE, 2010. 838–841.
15 Cotter SF. Weighted voting of sparse representation classifiers for facial expression recognition. Proceedings of the 2010 18th European Signal Processing Conference. Aalborg: IEEE, 2010. 1164–1168.
16 Li Y, Zeng JB, Shan SG, *et al*. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 2019, 28(5): 2439–2450. [doi: 10.1109/TIP.2018.2886767]
17 Rudovic O, Pantic M, Patras I. Coupled Gaussian processes for pose-invariant facial expression recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(6): 1357–1369. [doi: 10.1109/TPAMI.2012.233]
18 Lai YH, Lai SH. Emotion preserving representation learning via generative adversarial network for multi-view facial expression recognition. Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an: IEEE, 2018. 263–270.
19 Yovel G, Duchaine B. Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia. Journal of Cognitive Neuroscience, 2006, 18(4): 580–593. [doi: 10.1162/jocn.2006.18.4.580]
20 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.
21 Cheng JP, Dong L, Lapata M. Long short-term memory-networks for machine reading. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 551–561.
22 Long X, Gan C, De Melo G, *et al*. Attention clusters: Purely attention based local feature integration for video classification. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7834–7843.
23 Wang JF, Yuan Y, Yu G. Face attention network: An effective face detector for the occluded faces. arXiv: 1711.07246, 2017.
24 Yang JL, Ren PR, Zhang DQ, *et al*. Neural aggregation network for video face recognition. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5216–5224.
25 Yu KC, Salzmann M. Second-order convolutional neural networks. arXiv: 1703.06817, 2017.
26 Tuzel O, Porikli F, Meer P. Region covariance: A fast descriptor for detection and classification. Proceedings of the 9th European Conference on Computer Vision. Graz: Springer, 2006. 589–600.
27 Carreira J, Caseiro R, Batista J, *et al*. Semantic segmentation with second-order pooling. Proceedings of the 12th European Conference on Computer Vision. Florence: Springer, 2012. 430–443.
28 Acharya D, Huang ZW, Paudel DP, *et al*. Covariance pooling for facial expression recognition. Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018. 367–374.
29 Liu MY, Wang RP, Li SX, *et al*. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. Proceedings of the 16th International Conference on Multimodal Interaction. Istanbul: ACM, 2014. 494–501.
30 Huang ZW, Van Gool L. A Riemannian network for SPD matrix learning. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017. 2036–2042.
31 Wang K, Peng XJ, Yang JF, *et al*. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 2020, 29: 4057–4069. [doi: 10.1109/TIP.2019.2956143]
32 Li S, Deng WH. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing, 2019, 28(1): 356–370. [doi: 10.1109/TIP.2018.2868382]
33 Huang QH, Huang CQ, Wang XZ, *et al*. Facial expression recognition with grid-wise attention and visual transformer. Information Sciences, 2021, 580: 35–54. [doi: 10.1016/j.ins.2021.08.043]

(校对责编: 孙君艳)