【电子与信息科学 / Electronics and Information Science】

基于卷积-反卷积网络的正交人脸特征学习算法

孙文赞, 宋 昱, 陈昌盛

深圳市媒体信息内容安全重点实验室, 广东省智能信息处理重点实验室, 深圳大学电子与信息工程学院, 广东深圳 518060

摘 要: 身份特征与表情特征是人脸图像分析中的两组重要特征, 传统的有监督正交人脸特征学习 (supervised orthogonal facial feature learning, SOFFL) 算法虽然能够在给定表情和身份标签时学习这一对特征, 但因数据要求较高令其应用受限. 提出一种低数据要求的无监督正交人脸特征学习 (unsupervised orthogonal facial feature learning, UOFFL) 算法, 通过提取正交人脸特征的统一框架, 假设人脸图像空间中仅有身份和表情变化, 使用重构损失、分类损失和相关性最小化损失的组合, 采用深度卷积-反卷积神经网络, 从已对齐的人脸图像中联合学习, 提取身份和表情特征. 其中, 分类损失用于学习表情特征; 相关性最小化损失用于提高身份特征和表情特征之间的独立性; 重构损失用于确保两组特征组合的信息完整性. 在大规模合成人脸表情数据集 (large-scale synthesized facial expression dataset, LSFED) 和受限的 Radboud 人脸数据集 (Radboud faces dataset, RaFD) 上进行验证, 将所学身份特征空间中的欧氏距离用于人脸验证任务, 结果表明, 算法性能接近联合贝叶斯等有监督人脸识别方法. UOFFL 算法可在身份标签缺失的条件下, 仅使用表情特征学得身份特征. 相比改进前的 SOFFL 算法, 该方法缓解了对身份标签的依赖, 适用场合更广.

关键词: 人工智能; 计算机神经网络; 深度学习; 人脸表情识别; 人脸图像分析; 正交人脸特征; 重构损失; 分类损失; 相关性最小损失

中图分类号: TP311

文献标志码: A

doi: 10.3724/SP.J.1249.2020.05474

An orthogonal facial feature learning method based on convolutional-deconvolutional network

SUN Wenyun, SONG Yu, and CHEN Changsheng

Shenzhen Key Laboratory of Media Security, Guangdong Key Laboratory of Intelligent Information Processing,
College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, Guangdong Province, P. R. China

Abstract: In facial image analysis, the identity and expression are two important features which are related to face recognition and facial expression recognition tasks. The traditional supervised orthogonal facial feature learning method can learn both features jointly once the emotion and identity labels are given. Its application scope is limited by the high data requirement. In this paper, we propose a new method which is a united framework of joint facial feature learning with lower data requirement. It is based on the assumption that there are only two variations in the face space, i. e., the identity and the emotion. By combining reconstruction loss, classification loss, and correlation minimization loss, we use the deep convolutional-deconvolutional neural network to learn and extract identity and expression features from the aligned facial image. The classification loss learns the expression feature. The

Received: 2019-12-29; **Accepted:** 2020-02-07

Foundation: National Natural Science Foundation of China (61902250); Natural Science Foundation of Guangdong Province (2020A1515010563); Shenzhen Science and Technology Research Foundation for Basic Project (JCYJ20180305124550725)

Corresponding author: Lecture CHEN Changsheng. E-mail: cschen@szu.edu.cn

Citation: SUN Wenyun, SONG Yu, CHEN Changsheng. An orthogonal facial feature learning method based on convolutional-deconvolutional network [J]. Journal of Shenzhen University Science and Engineering, 2020, 37(5): 474-481. (in Chinese)

correlation minimization loss keeps the two features independent of each other. The reconstruction loss confirms the information completeness of the face. When the identity label is missing, the proposed method can learn the identity feature based on the expression labels. By this means, the new method is a kind of unsupervised facial feature learning. It relieves the limitation of the existing method, extends the application scope. The proposed method is evaluated on the large-scale synthesized facial expression dataset and the constrained Radboud face dataset (RaFD) dataset. The Euclid distance in the identity feature space is used for facial recognition task in which the performance of the proposed unsupervised facial feature learning is close to some supervision ones including the joint Bayesian face method. Our method relieves the requirement of identity label in the previous supervised version to get a wider application scope.

Key words: artificial intelligence; computer neural network; deep learning; facial expression recognition; facial image analysis; orthogonal facial feature; reconstruction loss; classification loss; correlation minimization loss

深度神经网络(deep neural networks, DNN)已在多种计算机视觉任务中获得成功应用,尤其是人脸表情识别与人脸识别任务.从数据中学习深度人脸特征是一种常见的做法,研究者通过设计不同的网络结构和训练方法获取具有不同性质的人脸特征.这些可学习的深度特征具有比传统图像特征更好的性质.例如,在人脸识别任务中,一个训练良好的身份特征对人脸的姿态、人脸表情、年龄和配饰具有不变性.类似地,在人脸表情识别任务中,表情特征对人脸的姿态、身份具有不变性.而在最新的域自适应算法中,训练良好的跨域特征可对域偏移具有不变性.

基于深度特征学习理论, SUN 等^[1]使用一对卷积-反卷积神经网络学习了身份特征与表情特征两组正交的人脸特征.其中,身份特征对表情的变化不变,而表情特征对身份变化不变.但是,该方法在特征学习过程中需要一对含表情的人脸和中性人脸,后者作为监督信息,用于定义同身份但不同表情的人脸图像的统一锚点.此外,该方法还需对每个身份分别记录7种表情(生气、恶心、恐惧、高兴、悲伤、惊讶和中性)的人脸图像.然而,主流的人脸识别数据集仅包含身份特征,并无关于表情的标签,少数的人脸表情识别数据集虽同时具有身份与表情标签,却并没有为每个身份提供对应的中性人脸.本研究为解决以上在联合学习过程中缺乏训练数据的问题,扩宽人脸身份与表情正交特征联合学习算法的应用范围,基于文献[1]方法,引入相关性最小化损失来缓解训练时对于中性人脸的依赖,通过构建一个卷积-反卷积神经网络,在已对齐人脸图像上提取身份与表情特征,并采用重构损失、分类损失和相关性最小化损失组合训练目标.本研究首次使用了表情标签学习身份特征,将新方

法记为无监督正交人脸特征学习(unsupervised orthogonal facial feature learning, UOFFL)算法,而将文献[1]方法记为有监督正交人脸特征学习(supervised orthogonal facial feature learning, SOFFL)算法.在大规模合成人脸表情数据集(large-scale synthesized facial expression dataset, LSFED)^[1]与受限的Radboud人脸数据集(Radboud faces dataset, RaFD)^[2]上的人脸特征学习结果表明, UOFFL算法的性能超越了一些无监督算法,且接近SOFFL等有监督算法.

1 基于深度特征分解的人脸特征学习

正交人脸特征学习方法将人脸特征分解为身份、姿态、表情和背景等部分,从任务角度来看, UOFFL算法与人脸特征分解^[1,3-8]有关;从损失与优化角度来看, UOFFL算法中的相关性最小化损失与基于协方差对齐和最小化的域自适应算法有关^[9-12],因为它们均通过优化来调整神经网络特征的分布,实现了特征的二阶统计量调节.

1.1 可分解深度特征学习

将深度人脸特征分解为身份、姿态、表情和背景等属性,一方面可从图像数据中学习、提取并分解不同的深度特征;另一方面,图像可从分解后的特征中重构和生成^[1,3-8].现有的可分解深度特征学习方法大多采用卷积神经网络提取特征,并使用反卷积神经网络重构和生成图像,也有采用基于对抗学习的生成模型^[3-6]. TRAN 等^[3]和 ZHANG 等^[4]分别提出可分解特征学习生成对抗网络和条件对抗自编码器.这两个网络将人脸分解为身份特征与姿态或年龄特征,并根据给定特征生成新的人脸图像. MA 等^[5]将行人图像分解为前景特征、背景特征与

姿态特征,并根据给定特征生成新的行人图像. BERTHELOT 等^[6]提出一种对抗受限自编码器插值的新型损失,可提高合成人脸的真实感和插值特征的语义连续性. 还有一些基于判别模型的图像重构方法^[1, 7-8]. ZHU 等^[7]使用非确定性神经元来采样随机的姿态特征,提出一种非确定性神经网络用于建模人脸、身份与姿态的关系. DOSOVITSKIY 等^[8]用反卷积神经网络学习属性到图像的函数映射. 本课题组 2018 年提出的 SOFFL 算法^[1]亦属于可分解深度特征学习范畴.

1.2 有监督正交人脸特征学习

SOFFL 算法从已对齐的人脸图像中提取身份特征与表情特征,训练样本为三元组 $(\mathbf{X}, \mathbf{y}, \mathbf{Z})$. 其中, $\mathbf{X}$ 为被提取特征的原始含表情人脸; $\mathbf{y}$ 为真实表情标签; $\mathbf{Z}$ 为对应于原始含表情人脸的中性人脸, $\mathbf{X}$ 和 $\mathbf{Z}$ 具有相同的身份和不同的表情. 在一些任务中,中性脸 $\mathbf{Z}$ 很难获取,因为主流的人脸识别数据集仅拥有身份标签,即使有些人脸表情识别数据集同时具有身份和表情标签,但也并非每个个体都有中性脸图像,训练数据要求过高是 SOFFL 算法的主要缺点.

1.3 基于二阶统计量调整的域自适应方法

二阶统计量调整是域自适应领域的主流方法之一^[9-12]. 其中,协方差对齐算法定义为

$$\min_{\mathbf{A}} \|\mathbf{A}^T \text{cov}(\mathbf{x}_s) \mathbf{A} - \text{cov}(\mathbf{x}_t)\|_F^2 \quad (1)$$

其中, $\mathbf{A}$ 为对齐源域和目标域特征的协方差的线性变换; $\mathbf{x}_s$ 与 $\mathbf{x}_t$ 分别为源域和目标域的样本域特征; 函数 $\text{cov}()$ 计算协方差矩阵. 文献[9]给出了式(1)的闭式解.

在神经网络中,协方差对齐问题常转化为

$$\min_{\theta} \|\text{cov}(f_{\theta}(\mathbf{x}_s)) - \text{cov}(f_{\theta}(\mathbf{x}_t))\|_F^2 \quad (2)$$

其中,神经网络 f_{θ} 受可训练参数 θ 控制; $f_{\theta}(\mathbf{x}_s)$ 和 $f_{\theta}(\mathbf{x}_t)$ 分别为待对齐的源域特征与目标特征的激活与目标激活,优化可使它们的分布在源域和目标域中一致. 式(2)常配合卷积神经网络的主要目标进行联合训练^[10-12],等价于最小化带有二次多项式核的最大均值差异(maximum mean discrepancy, MMD)^[11, 13-14]. 协方差对齐与核 MMD 均为用于神经网络的域自适应的主流算法.

另一种二阶统计量对齐方法为协方差最小化. BOUSMALIS 等^[11]提出基于二阶统计量的损失用于鼓励神经网络特征的各维度之间的分布差异:

$$\min_{\theta_c, \theta_p} \| [f_{\theta_c}(\mathbf{x})]^T f_{\theta_p}(\mathbf{x}) \|_F^2 \quad (3)$$

其中, $\mathbf{x}$ 为输入样本; f_{θ_c} 和 f_{θ_p} 分别是计算公共特征集和私有特征集的神经网络. 若 $f_{\theta_c}(\mathbf{x})$ 和 $f_{\theta_p}(\mathbf{x})$ 均已中心化,则式(3)等价于最小化两个特征集合之间的两两协方差

$$\min_{\theta_c, \theta_p} \|\text{cov}(f_{\theta_c}(\mathbf{x}), f_{\theta_p}(\mathbf{x}))\|_F^2 \quad (4)$$

这些域自适应方法初步验证了对神经网络激活的二阶统计量的调整的可行性.

1.4 无监督人脸验证

LIAO 等^[15]提出的无监督特征学习算法,可从一组姿态连续变化的人脸视频中学习人脸光照与姿态不变的特征. 该方法将相同身份的人脸分组在同一视频或集合中,分组被视为身份监督的一种变形. UOFFEL 算法则是从特征的互补性角度出发,因并未使用任何形式的身份监督,对训练数据要求更低.

2 无监督正交人脸特征学习

2.1 深度神经网络结构

为克服 SUFFL 算法训练数据要求高的缺点,本研究提出的 UOFFL 算法将针对 $\mathbf{Z}$ 的 L_2 判别损失替换为相关性最小化损失,来缓解对中性脸 $\mathbf{Z}$ 的依赖. 由于新损失是无监督的,故算法仅用二元组 $(\mathbf{X}, \mathbf{y})$ 进行训练,可视为 SOFFL 算法的无监督版本. 图 1 和图 2 分别为 SOFFL 算法和 UOFFL 算法的网络结构. 如图 1, SOFFL 算法所用的 L_2 损失与监督信号是图像,采用了一个 36 层的卷积-反卷积-卷积-反卷积网络进行原图-中性脸-重构图的变换,网络的第 18 层输出中性人脸图像 $\hat{\mathbf{Z}}$. UOFFL 算法则是将中性人脸图像 $\hat{\mathbf{Z}}$ 舍去,替换为一维的身份特征 $\mathbf{h}$,如图 2. 因此,SOFFL 算法实现了用一个 18 层的卷积-反卷积神经网络替代原先的 36 层网络,网络层数和可训练参数数量减半.

在图 2 中, l_{rec} 为重构损失; l_{cls} 为分类损失; l_{corr}

图 1 SOFFL 的网络结构^[1]

Fig. 1 The network architectures of SOFFL^[1]

图2 UOFFL 的网络结构

Fig. 2 The network architectures of UOFFL

为相关性最小化损失. 假设算法的已对齐人脸图像中仅存在身份和表情两种变化, 则在前向传播中, 卷积层将人脸图像编码为身份和表情特征, 而反卷积层从特征中重构输入的人脸图像.

记含有表情的人脸图像为 X , h 和 $\hat{y}_{\text{logit}}$ 分别为身份和表情特征, 二者组成卷积-反卷积网络的中间层激活为

$$(h, \hat{y}_{\text{logit}}) = f_{\theta_f}(X) \quad (5)$$

则网络的前向传播为

$$\hat{X} = g_{\theta_g}(h, \hat{y}_{\text{logit}}) \quad (6)$$

其中, $\hat{X}$ 为重构的人脸图像; $f_{\theta_f}()$ 和 $g_{\theta_g}()$ 为卷积层与反卷积层形式化定义函数, θ_f 和 θ_g 分别为其内含的可训练参数, 即神经网络的训练对象.

记预测概率 $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_C)^T$ 在对数空间中的点为 $\hat{y}_{\text{logit}} = (\hat{y}_{\text{logit}_1}, \hat{y}_{\text{logit}_2}, \dots, \hat{y}_{\text{logit}_C})^T$, 则通过标准 softmax 归一化函数计得预测概率为

$$\hat{y}_i = \frac{\exp(\hat{y}_{\text{logit}_i})}{\sum_{j=1}^C \exp(\hat{y}_{\text{logit}_j})}, i = 1, 2, \dots, C \quad (7)$$

本研究参考 VGG 网络的设计思路, 采用若干 3×3 卷积层后附一个下采样层组成的基础结构, 再由若干基础结构构成网络. 表 1 展示了 UOFFL 网络的层类型、层激活个数和层参数个数等细节, 整个网络是由一个卷积部分和一个反卷积部分组成, 其中卷积部分包含 6 个 VGG 基础结构, 将尺寸为 $64 \times 64 \times 1$ 像素的输入图像变换为 $1 \times 1 \times 519$ 像素的全局向量形式特征; 而反卷积部分也包含 6 个 VGG 基础结构, 其中后置下采样层替换为前置上采样层. 反卷积部分将 $1 \times 1 \times 519$ 像素的特征变换为 $64 \times 64 \times 1$ 像素的图像. 当输入尺寸、中间特征个数和输出尺寸确定时, 表 1 中的细节可按 VGG 网络的基本规则生成.

表1 UOFFL 的网络细节

Table 1 Details of the network for UOFFL

神经网络层	特征图尺寸/像素	参数数量/个
输入层	$64 \times 64 \times 1$	0
卷积层	$64 \times 64 \times 16$	192
卷积和下采样层	$32 \times 32 \times 16$	2 352
卷积层	$32 \times 32 \times 32$	4 704
卷积和下采样层	$16 \times 16 \times 32$	9 312
卷积层	$16 \times 16 \times 64$	18 624
卷积和下采样层	$8 \times 8 \times 64$	37 056
卷积和下采样层	$4 \times 4 \times 128$	74 112
卷积和下采样层	$2 \times 2 \times 256$	295 680
卷积和下采样层	$1 \times 1 \times (512 + 7)$	1 197 333
上采样和卷积层	$2 \times 2 \times 256$	1 196 544
上采样和卷积层	$4 \times 4 \times 128$	295 296
上采样和卷积层	$8 \times 8 \times 64$	73 920
上采样和卷积层	$16 \times 16 \times 64$	37 056
卷积层	$16 \times 16 \times 32$	18 528
上采样和卷积层	$32 \times 32 \times 32$	9 312
卷积层	$32 \times 32 \times 16$	4 656
上采样和卷积层	$64 \times 64 \times 16$	2 352
卷积层	$64 \times 64 \times 1$	147

使用批归一化(batch normalization, BN)、激活函数 $\tanh()$ 和平均池化下采样法来确保网络激活的是标准正态分布. 归一化激活分布不仅可加速训练还可简化相关性最小化损失中的皮尔逊相关系数(Pearson correlation coefficient, PCC)的计算. 每个卷积层使用 BN 归一化和 $\tanh()$ 激活. 整个网络共包括了 18 个卷积层, 不含任何全连接层, 网络共有约 328 万个可训练参数, 网络的规模远比一些常见网络小, 但已足够达到提取正交特征、重构原始人脸的要求.

2.2 损失函数

根据图 2, 网络在训练中使用了重构损失 l_{rec} 、分类损失 l_{cls} 与相关性最小化损失 l_{corr} . 首先讨论 l_{corr} , 从协方差对齐的角度出发, 需将式(2)对齐到 0 矩阵, 以缩小各特征维度之间的两两线性相关性. 从协方差最小化的角度(式(3)和式(4))出发, 需额外限制特征的中心, 即单一特征随机变量的期望值为 0. 如果采用传统的二阶统计量调整中的协

方差最小化方法^[10-13], 来降低特征 $\mathbf{h}$ 与 $\hat{\mathbf{y}}_{\text{logit}}$ 间相关性, 会导致平凡解, 即神经网络将学得较小的参数, 从而使激活较小, 进一步使激活函数 $\tanh(\cdot)$ 总是工作在其线性区域, 最终降低网络的非线性拟合能力, 减缓训练, 导致无法实现学习正交特征的初衷. 为此, 本研究将最小化协方差的方法改进为最小化特征 $\mathbf{h}$ 与特征 $\hat{\mathbf{y}}_{\text{logit}}$ 的两两皮尔逊相关系数. 随机变量 a 与 b 之间的皮尔逊相关系数为

$$\rho(a, b) = \frac{E[(a - E(a))(b - E(b))]}{\sigma(a)\sigma(b)} \quad (8)$$

其中, $E(\cdot)$ 和 $\sigma(\cdot)$ 分别为期望值与标准差函数. 因此, 两组随机向量 $\mathbf{a} = (a_1, a_2, \dots, a_m)^T$ 和 $\mathbf{b} = (b_1, b_2, \dots, b_n)^T$ 之间的皮尔逊相关矩阵可定义为 $\boldsymbol{\rho}(\mathbf{a}, \mathbf{b}) =$

$$\begin{bmatrix} \rho(a_1, b_1) & \rho(a_1, b_2) & \cdots & \rho(a_1, b_n) \\ \rho(a_2, b_1) & \rho(a_2, b_2) & \cdots & \rho(a_2, b_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(a_m, b_1) & \rho(a_m, b_2) & \cdots & \rho(a_m, b_n) \end{bmatrix} \quad (9)$$

PCC 是一种归一化的协方差, 值域为 $[-1, 1]$, 相比协方差矩阵, 它对随机变量的尺度具有不变性. 基于皮尔逊相关矩阵的相关性最小化损失为

$$l_{\text{corr}} = \|\boldsymbol{\rho}(\mathbf{h}, \hat{\mathbf{y}}_{\text{logit}})\|_F^2 \quad (10)$$

该损失在计算上等价于两两 PCC 的平方和, 即

$$l_{\text{corr}} = \sum_{i=1}^m \sum_{j=1}^n [\rho(\mathbf{h}_i, \hat{\mathbf{y}}_{\text{logit}_j})]^2 \quad (11)$$

其中, $\mathbf{h}_i$ 和 $\hat{\mathbf{y}}_{\text{logit}_j}$ 分别列举了 $\mathbf{h} = (h_1, h_2, \dots, h_m)^T$ 和 $\hat{\mathbf{y}}_{\text{logit}} = (\hat{y}_{\text{logit}_1}, \hat{y}_{\text{logit}_2}, \dots, \hat{y}_{\text{logit}_n})^T$ 中的所有特征. 显然, 最小化式(10)或式(11)可降低身份特征与表情特征之间的相关性.

计算 l_{corr} 需统计特征的均值与标准差, 在实践中, 由于深度神经网络往往采用随机梯度下降法训练, 在少量样本上估计均值与标准差会引入不稳定的噪声. 有学者采用滑动平均方法来获得稳定均值与标准差^[16], 而本研究采用另一种方法: 若随机变量 $\mathbf{h}$ 与 $\hat{\mathbf{y}}_{\text{logit}}$ 中的所有特征已是标准正态分布或某种已知分布, 则可避免从数据中估计分布参数. 因此, 将 BN 归一化后的特征 $\mathbf{h}$ 用于式(10)的相关性计算, 则 $E(\mathbf{h}) = 0$ 、 $\sigma(\mathbf{h}) = 1$. 用中心化后的真实独热码 $\mathbf{y} - 1/C$ 取代式(10)中的 $\hat{\mathbf{y}}_{\text{logit}}$, 则 $E(\mathbf{y} - 1/C) = 0$ 、 $\sigma(\mathbf{y} - 1/C) = \sqrt{(C-1)/C^2}$. 把上述常量 $E(\mathbf{h})$ 、 $E(\mathbf{y} - 1/C)$ 、 $\sigma(\mathbf{h})$ 和 $\sigma(\mathbf{y} - 1/C)$ 带入式(8)至式(10), 可实现高效且精确的前向和反向计算.

此外, 假设人脸图像 $\mathbf{X}$ 服从正态分布, 训练采

用一个 L_2 损失来确保两组特征的组合 $(\mathbf{h}, \hat{\mathbf{y}}_{\text{logit}})$ 信息完整, 即

$$l_{\text{rec}} = \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2 \quad (12)$$

用分类损失 l_{cls} 对比 C 维表情预测值 $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_C)^T$ 与真实表情标签 $\mathbf{y} = (y_1, y_2, \dots, y_C)^T$ 之间的差异

$$l_{\text{cls}} = - \sum_{i=1}^C [y_i \lg(\hat{y}_i) + (1 - y_i) \lg(1 - \hat{y}_i)] \quad (13)$$

假设人脸空间中仅有身份和表情两种变化, 则分类损失 l_{cls} 用于学习表情特征 $\mathbf{y}$; 重构损失 l_{rec} 用于确保信息完整性; 相关性最小化损失 l_{corr} 提升两组特征之间的独立性. 总体损失为三者的加权和

$$l_{\text{total}} = l_{\text{rec}} + \lambda_1 l_{\text{cls}} + \lambda_2 l_{\text{corr}} \quad (14)$$

其中, 非负权重 λ_1 和 λ_2 用于平衡 3 个分量的重要性.

3 实验与结果分析

3.1 数据集与预处理

采用 LSFED^[1] 与 RaFD^[2] 数据库验证 UOFFL 算法的性能. 由于 LSFED 数据库中的图像由软件合成, 不含噪声且对齐精准, 导致所有对比算法的性能都趋于饱和, 无法区别算法优劣, 故提升难度以模拟复杂的真实世界环境, 给数据集加入 SNR = 20 dB 的高斯噪声, 并将加噪的数据集记为 LSFED-G. 根据文献[15]方法构造包含高斯噪声和随机相似变换的数据集, 记为 LSFED-GS; 构造包含高斯噪声、随机相似变换和随机背景的数据集, 记为 LSFED-GSB.

RaFD 是一个在受限环境下采集的小规模人脸表情数据集, 仅包含 67 个个体, 每个个体有 8 种表情、5 种姿态和 3 个眼睛注视方向信息. 为保持数据集的一致性, 本实验仅使用眼睛直视的正面脸图像, 并丢弃轻蔑表情的图像, 最终剩下 469 张人脸图像. 在处理中, 首先使用基于方向梯度直方图 (histogram of oriented gradient, HOG) 特征与支持向量机 (support vector machine, SVM) 的人脸检测器检测人脸包围盒; 然后检测 68 个人脸特征点^[17], 估计人脸形状与姿态^[18]; 最后将人脸图像对齐到预定义的三维人脸几何体上, 再渲染和裁剪出尺寸为 64×64 像素的人脸图像. 图 3 为部分预处理后的人脸.

两个数据集大致按照 8:2 的比例划分为训练集和测试集, 且其中的人脸所属的身份不重合.

图3 部分预处理后的 LSFED 与 RaFD 数据集人脸

Fig. 3 Some preprocessed faces in LSFED dataset and RaFD dataset

3.2 训练模型

采用文献[19]方法初始化网络的卷积核, 偏置被初始化为 0, BN 中的均值和标准差的滑动平均初始值分别设为 0 和 1, 采用自适应矩估计 (adaptive moment estimation, ADAM) 优化器训练网络, 优化器参数设置为 $\alpha = 0.001$ 、 $\beta_1 = 0.9$ 、 $\beta_2 = 0.999$, $\varepsilon = 1 \times 10^{-8}$. 使用随机梯度下降方式最小化总体损失. 训练集先被随机打乱顺序, 每次迭代依次送入 100 个训练样本, 每个训练样本均参与训练 100 次后终止训练. 为评价 3 个损失分量对识别结果的影响,

采用不同的 λ_1 和 λ_2 值, 当 $\lambda_1 = \lambda_2 = 0$ 时对应的损失分量不发挥作用.

3.3 无监督人脸验证对照实验结果分析

基于已训练的网络, 使用人脸验证任务评价学得的身份特征 h 的性能. 随机抽取 1 000 个正样本对 (相同身份但不同表情) 和 1 000 个负样本对 (不同身份但相同表情), 计算样本对在身份特征空间中的欧氏距离, 并选取合适的阈值. 若两个人脸的距离大于阈值则判别为不同身份; 若小于阈值则判别为相同身份. 选取的接收者操作特征曲线上面积 (area under the receiver operating characteristic curve, AUC) 和等错误率 (equal error rate, EER) 评价指标与阈值无关, 在应用中可选取训练集上距离的中值作为人脸验证的阈值. 表 2 展示了不同数据集在不同空间欧氏距离中采用无监督人脸识别所得 AUC 和 EER 值. 由表 2 可见, UOFFL 算法在 LSFED、RaFD 和加噪的数据集上工作良好, 在相对干净的 LSFED、LSFED-G 与 RaFD 数据集上, 取 $\lambda_1 = 1$, $\lambda_2 = 1$ 时可获得较好的结果, 而在有严重噪声的 LSFED-GS 与 LSFED-GSB 数据集上, 取 $\lambda_1 = 1$, $\lambda_2 = 10$, 适当提高相关性最小化损失的作用, 提升身份特征与表情特征之间的独立性, 可获得较好的身份特征, 进而获得更好的人脸验证性能.

表2 基于不同空间欧氏距离的无监督人脸识别性能结果¹⁾

Table 2 Unsupervised face recognition results based on Euclidean distance in different spaces

数据集	X 空间中的 AUC/EER 值	h 空间中的 AUC/EER 值				
		$\lambda_1 = 0,$ $\lambda_2 = 0$	$\lambda_1 = 1,$ $\lambda_2 = 0$	$\lambda_1 = 1,$ $\lambda_2 = 1$	$\lambda_1 = 1,$ $\lambda_2 = 10$	$\lambda_1 = 1,$ $\lambda_2 = 100$
LSFED	0.985/6.4	0.993/4.8	0.997/2.5	1.000/0	0.999/0.8	0.971/8.3
LSFED-G	0.985/6.5	0.976/9.8	0.996/3.0	1.000/0	1.000/0.1	0.984/6.9
LSFED-GS	0.781/30.1	0.787/29.7	0.811/27.1	0.761/31.7	0.920/16.8	0.835/27.6
LSFED-GSB	0.613/42.2	0.695/35.4	0.711/34.5	0.735/33.7	0.768/31.2	0.562/47.0
RaFD	0.814/27.2	0.943/14.6	0.952/12.9	0.974/8.3	0.962/10.6	0.845/23.4

¹⁾ 灰底数值为该指标上取得的最优结果

表 2 第 2 列 X 空间中的 AUC 和 EER 值是将原始图像拉直为向量后, 计算欧氏距离所得的 AUC 和 EER 指标结果. 第 3 ~ 7 列为不同 λ_1 和 λ_2 取值下 UOFFL 算法的 AUC 和 EER 指标结果, 使用特征学习可以获得比原始图像上更好的结果. 对比第 3 ~ 6 列可以发现, 启用相关性最小化损失, 并选取合适的权重可分别在 LSFED、LSFED-G、LSFED-GS、LSFED-GSB 和 RaFD 数据集上获得 0.002 7、0.004 1、0.108 9、0.057 5 与 0.021 8 的 AUC 指标

提升, 故 UOFFL 算法的相关性最小化是解决无监督人脸验证问题的关键. 选择合适的 λ_1 和 λ_2 可进一步提升性能, 在大多数情况下, 取 $\lambda_1 = \lambda_2 = 1$ 较好, 当人脸图像中噪声交大时, 可尝试取 $\lambda_1 = 1$, $\lambda_2 = 10$.

3.4 有监督和无监督人脸验证的对比

表 3 对比了 UOFFL 算法与其他 7 种人脸验证算法的 AUC 指标值. 其中, LBP + PCA + LDA^[20]、AlexNet 预训练 + 微调^[21]、两层神经网络、LBP +

PCA + 联合贝叶斯^[22]和 SOFFL 算法^[1]因训练中使用身份标签属有监督学习; UOFFL 算法、原图上的欧氏距离方法、PCA 主成分上的欧氏距离方法和文献[14]算法是未使用身份标签的无监督学习. UOFFL 算法在相对干净的 LSFED、LSFED-G 与 RaFD 数据集上的表现超过或接近有监督学习方法; 在噪声较大的 LSFED-GS 数据集上, UOFFL 算法在

无监督算法中的表现仍保持最好; 而在含有随机背景的 LSFED-GSB 数据集上, UOFFL 算法的性能不佳, 这是因为随机背景作为表情特征的互补亦被编码在身份特征中, 影响人脸验证效果. LSFED-GSB 数据集违背了 UOFFL 算法人脸空间中仅有身份和表情两种变化的假设, 故需根据实际情况选择算法.

表 3 有监督和无监督人脸验证方法的性能对比¹⁾

Table 3 Comparisons among supervised and unsupervised facial verification methods

类 型	算 法	不同数据集上的人脸验证 AUC 指标				
		LSFED	LSFED-G	LSFED-GS	LSFED-GSB	RaFD
有监督	LBP + PCA + LDA ^[20]	0.998	0.983	0.918	0.941	0.928
	AlexNet 预训练 + 微调 ^[21]	0.994	0.991	0.968	0.932	0.928
	两层神经网络	1.000	0.998	0.978	0.970	0.925
	LBP + PCA + 联合贝叶斯 ^[22]	1.000	1.000	1.000	0.985	0.972
	SOFFL	1.000	1.000	0.999	0.998	0.962
无监督	原图上的欧氏距离	0.985	0.985	0.781	0.613	0.814
	PCA 主成分上的欧氏距离	0.993	0.995	0.860	0.808	0.836
	旋转与光照不变特征学习 ^[14]	1.000	1.000	0.904	0.751	0.897
	UOFFL	1.000	1.000	0.920	0.768	0.974

¹⁾ 灰底数值为在该指标上取得的最优结果

UOFFL 算法与 SOFFL 算法关系紧密, 不同于主流的已在人脸基准上获得较好的性能的人脸验证算法, SOFFL 算法需要包含表情人脸和同身份中性表情人脸图像对作为训练数据, 其训练样本为三元组 (X, y, Z) . 由于对数据要求过高, SOFFL 方法无法在野外标签人脸 (labeled faces in the wild, LFW) 数据库等主流数据集上训练测试, 本研究将其训练数据简化为二元组 (X, y) , 导致 UOFFL 算法在 LSFED-GS 与 LSFED-GSB 数据集上的性能略低于 SOFFL 方法. 然而, 在小样本 RaFD 数据集上, UOFFL 算法的确比 SOFFL 算法略有提升 (由 0.962 提升至 0.974), 这是因为 SOFFL 算法需要学习一个复杂的图像到图像双向映射, 这在小样本数据上会比较困难, 而 UOFFL 算法改为学习像素图像到 519 维特征的双向映射, 避免了该问题.

尽管有监督方法比无监督方法因使用更多的数据而获得了更好的性能, 但 UOFFL 算法提出的相关性最小化损失, 能够缓解标签缺失的劣势, 缩小无监督和有监督方法的性能差距. 总体上讲, UOFFL 算法性能优于无监督的文献[14]和有监督的 SPFFL 算法, 接近有监督的联合贝叶斯人脸识别算法.

结 语

提出一种基于卷积-反卷积网络的正交人脸特征学习 UOFFL 算法, 使用 3 个损失训练网络: 分类损失用于学习表情特征, 重构损失用于确保特征中信息的完整性, 相关性最小化损失用于提高身份特征与表情特征之间的独立性, 最后, 在合成人脸和真实人脸数据集上进行实验验证. 下一步, 计划将方法拓展到非受限人脸和跨库或跨域问题中. 例如, 在野外静态面部表情 (static facial expressions in the wild, SFEW) 数据集上训练, 并在 LFW 数据集上测试, 或将 LSFED 数据集上学得的知识迁移到非受限人脸.

基金项目: 国家自然科学基金资助项目 (61902250); 广东省自然科学基金资助项目 (2020A1515010563); 深圳市科技计划基础研究资助项目 (JCYJ20180305124550725)

作者简介: 孙文赞 (1987—), 深圳大学博士后研究人员. 研究方向: 深度学习与人脸图像分析. E-mail: wenyunsun@szu.edu.cn

引 文: 孙文赞, 宋 昱, 陈昌盛. 基于卷积-反卷积网络的正交人脸特征学习算法[J]. 深圳大学学报理工版, 2020, 37(5): 474-481.

参考文献 / References:

- [1] SUN Wenyun, ZHAO Haitao, JIN Zhong. A complementary facial representation extracting method based on deep learning [J]. *Neurocomputing*, 2018, 306 (6): 246-259.
- [2] LANGNER O, DOTSCH R, BIJLSTRA G, et al. Presentation and validation of the Radboud faces database [J]. *Cognition and Emotion*, 2010, 24(8): 1377-1388.
- [3] TRAN L, YIN Xi, LIU Xiaoming. Disentangled representation learning gan for pose-invariant face recognition [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE, 2017: 1415-1424.
- [4] ZHANG Zhifei, SONG Yang, QI Hairong. Age progression/regression by conditional adversarial autoencoder [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE, 2017: 5810-5818.
- [5] MA Liqian, SUN Qianru, GEORGOULIS S, et al. Disentangled person image generation [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE, 2018: 99-108.
- [6] BERTHELOT D, RAFFEL C, ROY A, et al. Understanding and improving interpolation in autoencoders via an adversarial regularizer [EB/OL]. (2018-07-18) [2018-07-23]. <https://arxiv.org/abs/1807.07543>, 2018.
- [7] ZHU Zhenyao, LUO Ping, WANG Xiaogang, et al. Multi-view perceptron: a deep model for learning face identity and view representations [C]// *Proceedings of the 27th Conference on Neural Information Processing Systems*. Montreal, Canada: MIT Press, 2014, 1: 217-225.
- [8] DOSOVITSKIY A, SPRINGENBERG J T, BROX T. Learning to generate chairs with convolutional neural networks [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA: IEEE, 2015: 1538-1546.
- [9] SUN Baochen, FENG Jiashi, SAENKO K. Return of frustratingly easy domain adaptation [C]// *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. Phoenix, USA: AII Press, 2016: 2058-2065.
- [10] SUN Baochen, SAENKO K. Deep coral: correlation alignment for deep domain adaptation [C]// *Proceedings of the European Conference on Computer Vision*. [S. l.]: Springer, 2016: 443-450.
- [11] BOUSMALIS K, TRIGEORGIS G, SILBERMAN N, et al. Domain separation networks [C]// *Advances in Neural Information Processing Systems*. New York, USA: Curran Associates Inc., 2016: 343-351.
- [12] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA: IEEE, 2016: 2414-2423.
- [13] ROZANTSEV A, SALZMANN M, FUA P. Beyond sharing weights for deep domain adaptation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 41(4): 801-814.
- [14] LI Yanghao, WANG Naiyan, LIU Jiaying, et al. Demystifying neural style transfer [C]// *Proceedings of the International Joint Conference on Artificial Intelligence*. Palo Alto, USA: AAAI Press, 2017: 2230-2236.
- [15] LIAO Qianli, LEIBO Joel Z, POGGIO Tomaso. Learning invariant representations and applications to face verification [C]// *Advances in Neural Information Processing Systems*. Lake Tahoe, USA: MIT Press, 2013: 3057-3065.
- [16] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [EB/OL]. (2015-02-11) [2015-03-02]. <https://arxiv.org/abs/1502.03167>, 2015.
- [17] KAZEMI V, SULLIVAN J. One millisecond face alignment with an ensemble of regression trees [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, USA: IEEE, 2014: 1867-1874.
- [18] HASSNER T, HAREL S, PAZ E, et al. Effective face frontalization in unconstrained images [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA: IEEE, 2015: 4295-4304.
- [19] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: surpassing human-level performance on imageNet classification [C]// *Proceedings of the IEEE International Conference on Computer Vision*. Santiago, Chile: IEEE, 2015: 1026-1034.
- [20] CHEN Dong, CAO Xudong, WEN Fang, et al. Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification [C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Portland, USA: IEEE, 2013: 3025-3032.
- [21] DONAHUE J, JIA Yangqing, VINYALS O, et al. DeCAF: a deep convolutional activation feature for generic visual recognition [C]// *Proceedings of the International Conference on Machine Learning*. Beijing: JMLR.org, 2014, 32: 647-655.
- [22] CHEN Dong, CAO Xudong, WANG Liwei, et al. Bayesian face revisited: a joint formulation [C]// *Proceedings of the 12th European Conference on Computer Vision*. Berlin: Springer, 2012: 566-579.

【中文责编: 英 子; 英文责编: 木 柯】

<http://journal.szu.edu.cn>