有遮挡人脸识别综述:从子空间回归到深度学习

李小薪 梁荣华
(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 有遮挡人脸识别是面向现实的人脸识别系统需要重点解决的问题,其困难性主要体现在由遮挡所引发的特征损失、对准误差和局部混叠等方面.该文从鲁棒分类器的设计和鲁棒特征提取两方面回顾了现有的方法.充分利用人脸图像和遮挡自身所固有的结构来表示、抑制或消除遮挡或由遮挡引发的误差是目前设计鲁棒分类器的关键思路.从子空间回归的角度回顾了主流的线性回归分类器处理遮挡问题的一般方法:协同表示、遮挡的字典表示及遮挡字典的学习和压缩技术;从结构化误差编码的角度回顾了基于人脸图像低秩结构的误差编码方法和将遮挡的空间结构嵌入重构误差的编码方法;从噪声抑制和遮挡检测两方面回顾了现有的迭代重权误差编码方法.文中强调特征提取对于解决有遮挡人脸识别问题的重要性,总结了鲁棒特征提取的基本要素,深入分析了以图像梯度方向和韦伯脸为代表的"浅层"特征所引发的零和差异现象、以PCANet为代表的将卷积神经网络与经典的"特征图-模式图-柱状图"特征提取框架相结合的编码原理,以及以DeepID为代表的卷积神经网络所生成的"深度"特征所具有的遮挡不变性及其所蕴含的编码准则.在Extended Yale B、AR和LFW等三个基准数据库上对现有方法的有效性进行了大规模测试,指出了现有方法的适用面及局限性.最后指出了有遮挡人脸识别给计算机视觉带来的挑战、现有方法在优化算法和特征提取方面存在的主要问题以及未来利用卷积神经网络处理遮挡问题需重点考虑的问题.

关键词 人脸识别;子空间回归;误差编码;遮挡字典;鲁棒特征;图像分解;深度学习
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2018.00177

A Review for Face Recognition with Occlusion: From Subspace Regression to Deep Learning

LI Xiao-Xin LIANG Rong-Hua
(*College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou* 310023)

Abstract Real world face recognition system has to struggle with facial occlusion problem. The practical facial occlusions can be roughly divided into three categories: the occlusion caused by illumination changes, the disguise incurred by material objects, and the self-occlusion rising from pose variations. The difficulties of face recognition with occlusion mainly lie in the occlusion inducing problems, such as feature missing, alignment error and local feature alias. We review the extant methods against facial occlusion from two aspects: the robust classification and the robust feature extraction. The main ideas of designing robust classifiers is to represent, weaken or eliminate facial occlusion *per se* or the errors caused by facial occlusion by fully considering the inherent structures of the facial images and the facial occlusions. From the view of subspace regression, we review the main methods of the popular linear robust classifiers against occlusion, including collaborative representation, occlusion coding with the well designed or learned occlusion dictionary, and occlusion dictionary learning and compression; from the view of structured error coding, we review the error coding methods based on the low-rank structure of facial images, and the error

收稿日期:2016-08-11;在线出版日期:2017-06-01.本课题得到国家自然科学基金(61402411,61379017,61672464)、浙江省自然科学基金(LY18F020031,LY17F020021)资助.**李小薪**,男,1980年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为计算机视觉、模式识别和图像处理.E-mail:mordekai@zjut.edu.cn.**梁荣华**(通信作者),男,1974年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为图像处理与计算机视觉.E-mail:rhliang@zjut.edu.cn.

coding methods with the occlusion spatial structures embedding using the structure inducing norms, such as the nuclear norm; from the view of noise weighting and occlusion detection, we review the extant iteratively reweighting error coding methods, which fully considers the statistical or spatial relationships between facial occlusion and the error incurred by facial occlusion. We stress that feature extractions are very important and critical in handling facial occlusion, and summarize four basic elements from the existing robust feature extraction methods, which are image decomposition, high-order semantic feature extraction, spatial locality and multi-hierarchy handling. For the shallow features, such as image gradient orientations and Weberfaces, we deeply analyze their robustness mechanism, i. e. , the phenomena of the underlying zero-sum differences incurred by these features. We then go deep into the coding principle of PCANet, which brings the deep filtering thought of the convolutional neural networks (CNN) into the classical feature extraction framework, which consists of feature map generation, pattern maps computing and local histogram producing. For the deep features produced by CNN, such as DeepID, VGG and LCNN, we illustrate their occlusion-invariant property and elaborate their coding criteria. We validate the efficacy of the extant methods in dealing with facial occlusions on three benchmark face databases, i. e. , the Extended Yale B, the AR and the LFW, and indicate their applicable problems and limitations. We finally point out the main challenges brought by face recognition with occlusion in the field of computer vision and pattern recognition, and the problems of the existing methods in their solving optimization algorithms and their feature extraction procedures. We also indicate the main problems that CNNs should consider in dealing with facial occlusions in the future: the transfer learning from the unconstrained datasets to the constrained ones, the training challenges with only small training sample size, the usage of the occlusion priori widely adopted in the classical error coding methods and feature extraction methods, and the relationships between the sparsity of the CNN models with their robustness.

Keywords face recognition; subspace regression; error coding; occlusion dictionary; robust features; image decomposition; deep learning

1 引 言

在过去的二十年中,人脸识别技术尽管已经取得了长足的进展[1],然而,面向现实的人脸识别系统仍然面临着诸多挑战.这些挑战主要来自于人脸图像中可能存在的多种多样的难以预测的变化,如表情、姿势、光照、分辨率和遮挡等.这些变化可能导致训练图像和待识别图像之间发生严重的数据偏移[2].本文重点关注有遮挡人脸识别问题.

在实际生活中,引发面部遮挡的原因可能有很多,如图 1 所示,归纳起来主要有以下 3 种:(1)由不均匀的或极度强烈的外部光照所引起的"光线遮挡";(2)由覆盖于人脸表面的外界物体,如手、帽子、太阳镜、围巾等所引起的"实物遮挡";(3)由人脸自身的姿势变化所引起的"自遮挡"[4].另外,现实世界中也可能存在更为复杂的混合遮挡,如图 1(d)和(e)所示.Ding 等人[5]回顾了近十年来人脸识别问题中姿势变化问题的解决方案.本文假定对姿势变化的处理可以独立于对遮挡的处理,重点关注光线遮挡和实物遮挡.

(a) (b) (c) (d) (e)

图 1 现实世界中的各种遮挡((a)光线遮挡;(b)实物遮挡;(c)由姿势变化引发的自遮挡;(d)、(e)混合遮挡)

近年来,继 Wright 等人[6]的工作之后,有遮挡人脸识别问题引起了广泛的关注[7-22].已有的工作主要可以分为子空间回归、鲁棒误差编码与鲁棒特征提取等三类.子空间回归方法[6,13,20,22-26]将来自于不同类别的人脸图像划分为不同的子空间,并且为遮挡建立独立的遮挡子空间,认为有遮挡人脸图像

$y\in\mathbb{R}^m$是原始的不含遮挡的人脸图像$y^0\in\mathbb{R}^m$与遮挡$o\in\mathbb{R}^m$的叠加：$y=y^0+o$，从而把有遮挡人脸图像y的识别问题视作将y^0和o各自回归到它们所属的子空间的问题. 子空间回归方法的主要困难在于遮挡子空间的构建(或遮挡字典的设计).

鲁棒误差编码方法认为有遮挡图像y是原始的不含遮挡的人脸图像y^0与由遮挡引发的误差$e\in\mathbb{R}^m$的合成体，主要包括“加法模型”和“乘法模型”. 加法模型又称为生成模型[11,26-29]，将有遮挡图像y看作原始图像y^0和误差e的加性合成体：$y=y^0+e$，着重考虑如何将误差e从y中分离出来(类似于“盲源分离”[30]，子空间回归方法在本质上也可归结为此类模型)，往往需要对误差e进行复杂的描述.

乘法模型又称为判别模型[9,10,12,14,19]，将有遮挡图像y看作有遮挡部分和无遮挡部分的拼接，认为只有其无遮挡部分是可以精确重构的：$w\odot y=w\odot(y^0+e)$($w\in[0,1]^m$或$w\in\{0,1\}^m$)，着重考虑如何分离其有遮挡区域和无遮挡区域，往往需要对误差权重w进行复杂的描述. He 等人[31]将鲁棒误差编码的加法模型和乘法模型统一纳入基于半二次型的鲁棒误差编码框架①，并在该框架下研究两者的关系：可以分别用半二次型的加法形式和乘法形式对二者建模.

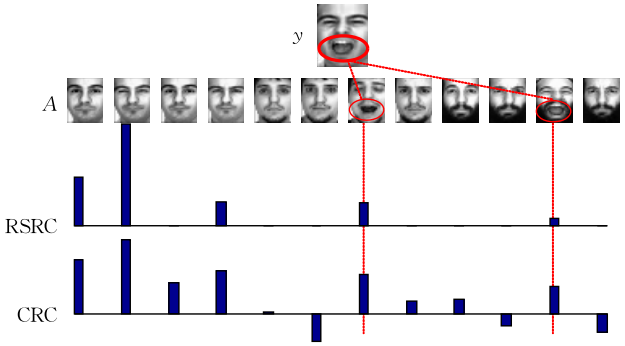
子空间回归方法和鲁棒误差编码方法同属于鲁棒分类器的范畴，它们都具有特征选择和特征编码的功能. Wright 等人[6]认为特征提取对于鲁棒的稀疏分类器而言不再重要，并且指出：就有遮挡图像而言，没有任何特征比原始图像本身更为冗余、鲁棒、局部和富含信息，全局性的特征提取技术(如 Eigenfaces 和 Fisherfaces 等)会将局部遮挡扩散到新特征的全局范围，而局部特征提取技术(如 LBP、Gabor 变换、ICA、LNMF 等)会使得遮挡特征在局部范围内扩散. 然而，近期大量的研究工作[13,17-18,25-26,32-34]表明局部特征提取对有遮挡人脸图像识别仍然是至关重要的：对于 PCANet 等“深度”特征，Chan 等人[17]的实验表明，即使只用最简单的最近邻分类器也能达到非常优秀的识别性能；对于图像梯度方向等“浅层”特征，本文的大量实验表明将其嵌入鲁棒分类器，也会达到甚至部分超过基于深度特征的识别效果.

尽管已经取得了长足的进展，有遮挡人脸识别仍然面临着巨大的挑战：(1) 现有的方法仍然不能完全有效地排除遮挡的影响. 本文图 15(7.2 节)针对有模拟遮挡的人脸识别实验表明：如果已知遮挡的位置，并且完全排除了遮挡的影响，那么，即使测试图像包含了较大面积的遮挡(遮挡比例达到 80% 和 90%)，即便只使用简单的“浅层”特征和最近子空间分类器，只要训练样本足够丰富，也能达到近于 100% 的识别率(Li 等人[9]也给出了类似的结论)；而在遮挡位置未知的情况下，即便有足够丰富的训练样本并且使用最前沿的深度特征(如 PCANet[17])，大面积的遮挡仍然会导致识别率的急剧下降. 这表明遮挡本身的存在比遮挡所造成的特征损失对识别性能的影响更为严重；(2) 视觉经验告诉我们，人眼在观察一幅有遮挡的图像时很容易感知到遮挡的位置，而且不需要大规模的训练，然而，对于计算机视觉而言，遮挡感知仍然是困难的，甚至判断一幅图像中是否存在遮挡都是困难的.

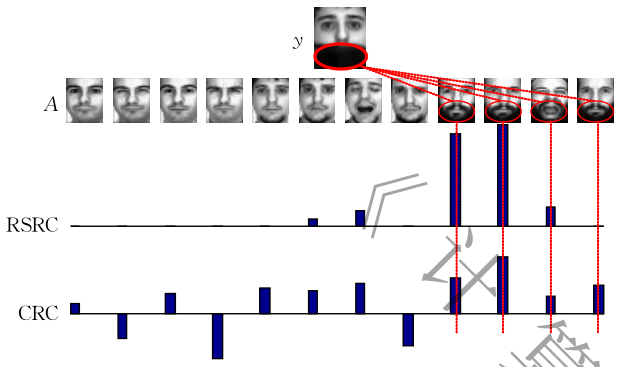
如果遮挡不能完全排除，而且遮挡所造成的特征损失对识别性能的影响又不是至关重要的(假定有足够丰富的训练样本)，那么，影响有遮挡人脸识别性能的关键因素是什么？Ekenel 等人[35]指出：由遮挡所引发的对准误差是导致有遮挡图像识别性能下降的关键因素. 本文 7.2 节的实验也验证了这一点. 然而，对于 7.2 节的实验而言，由于所添加的是模拟遮挡，对准误差是不存在的. 通过实验可以发现，“局部混叠”或“局部相干”是导致有遮挡人脸识别困难的另一个主导因素. 所谓“局部混叠”或“局部相干”是指：待识别图像中遮挡所形成的局部特征可能会与训练集中其它类别的图像所固有的局部特征非常相似(具有很强的相干性)，从而使得分类器“混淆”了该类别的图像和真实类别的图像. 图 2 针对主流的鲁棒分类器 RSRC[6]和 CRC[24]，选取 AR 人脸数据库[36]中的有表情变化和围巾遮挡的测试图像说明了这一现象：遮挡比一般的变化(如表情变化)更容易引起混叠.

鉴于上述有遮挡人脸识别中存在的关键问题和挑战，我们有必要深入了解现有的有遮挡人脸识别方法. 本文从鲁棒分类器的设计和鲁棒特征提取两方面对现有的方法进行了回顾和分析，提出未来解决有遮挡人脸识别问题的四个思路：(1) 寻求能够表达遮挡的固有结构的误差编码方法是设计鲁棒分类器的关键任务；(2) 特征提取仍然是解决有遮挡人脸识别问题的关键，而且未来特征提取的主导问

① He 等人[31]将加法模型和乘法模型分别称为误差校正模型和误差检测模型，但“误差校正”的概念过于宽泛，在本质上涵盖了误差检测，如 Zhou 等人[12]的 SEC-MRF 模型本身就称为“误差校正模型”，但同时也是一种误差检测模型，所以本文采用加法模型和乘法模型的说法.



(a) 待识别图像中“张嘴”的表情同时存在于第2个人和第3个人的样本中，但并未引发鲁棒分类器的混叠



(b) 与待识别图像中的围巾遮挡类似的“黑色胡子”存在于第3个人的样本中，导致了鲁棒分类器的混叠

图 2 遮挡比表情变化更容易引发鲁棒分类器(RSRC/CRC)的“局部混叠”(柱状图的纵坐标表示由 RSRC/CRC 算得的回归系数的大小)

题之一是如何通过滤波器的学习，直接消除“遮挡”的影响而不会将其扩散到更大范围；(3) 基于小样本训练具有局部连续结构的异常特征的检测网络. 对具有局部连续结构的异常特征而言，人类视觉系统只要基于小样本的训练就可以在新的目标图像上准确捕捉到异常特征的位置，这与以卷积神经网络为代表的深度学习方法所基于的大规模训练机制在本质上是不同的(基于更小规模的样本训练深度网络也是刚刚兴起的研究课题[37,38])；(4) 将鲁棒特征提取技术嵌入已有的鲁棒分类器中，以充分利用鲁棒特征的性能和发挥鲁棒分类器的误差表示和误差检测的能力.

本文第 2 节从子空间回归的角度回顾主流的线性回归分类器处理遮挡问题的一般方法；第 3 节从人脸图像自身所固有的结构和遮挡所固有的结构回顾现有的结构化误差编码方法；第 4 节基于人类视觉的局部相似匹配原理和遮挡的空间先验结构回顾现有的迭代重权误差编码方法；第 5 节回顾主流的鲁棒特征提取技术；第 7 节比较和分析本文所讨论的主要分类方法和特征提取技术在 Extended Yale B、AR 和 LFW 等人脸数据库上的识别性能；第 8 节总结现有方法仍然存在的问题与未来研究的重点. 另外，考虑到一般的情形，本文假设：训练集不包含并且也不知道任何遮挡信息，只知道遮挡在图像空间中的分布是局部的和连续的.

2 子空间回归

子空间回归方法基于一个简单的假设：某一类别的样本来自于同类样本所张成的低维子空间，识别问题就可归结为使得待识别样本回归到它所属的子空间的问题. 假设有来自于 K 个人的 n 个训练样本构成的字典 $A=[A^1,A^2,\cdots,A^K]\in\mathbb{R}^{m\times n}$，其中，$A^k=[a_1^k,a_2^k,\cdots,a_{n_k}^k]\in\mathbb{R}^{m\times n_k}$ 为第 k 个人的 n_k 个样本的集合，$n=\sum_{k=1}^{K}n_k$，每个样本 $a_i^k\in\mathbb{R}^m$ 也称为训练字典 A 的“原子”. 那么，来自于第 k 个人的 n_k 个样本 A^k 就张成了第 k 个人的低维子空间.

人脸图像之间天然的强相关性使得各人脸子空间在高维图像空间中的分布十分紧凑[23]，如何去除不同子空间中的人脸图像的相干性并增强同一子空间中人脸图像之间的相干性是子空间回归方法需要重点考虑的问题[39]. 对于有遮挡人脸识别问题，还要考虑到遮挡子空间如何形成，以及如何去除人脸子空间和遮挡子空间的相干性等问题.

2.1 稀疏编码与协同表示

由 Wright 等人[6]于 2009 年提出的基于稀疏表示分类方法(Sparse Representation-based Classification，SRC)和 Zhang 等人[24]于 2011 年提出的基于协同表示的分类(Collaborative Representation-based Classification，CRC)是两个最具有代表性的子空间回归方法.

SRC 模型[6]基于压缩传感理论[40]，通过最小 ℓ_1 范数回归从相关性较强的各人脸子空间中寻找能够最有力表达待识别图像 y 的字典原子：

$$\min_x \|x\|_1 \quad \text{s.t.} \quad y=Ax \tag{1}$$

其中，$x\in\mathbb{R}^n$ 为重构系数(也称为回归系数或编码系数). 从而使得重构系数 x 的非零元素尽可能落在 y 所属的子空间中. Wright 等人[6]指出 SRC 的重构能力主要依赖于待识别图像 y 的维数 m，而不依赖于 y 的具体特征. 为了有效地处理有遮挡人脸识别问题，Wright 等人[6]又进一步提出了鲁棒的 SRC(Robust SRC，RSRC)模型：

$$\min_{x,e}\|e\|_1+\|x\|_1, \quad \text{s.t.} \quad y=[AA_o]\begin{bmatrix}x\\e\end{bmatrix} \tag{2}$$

其中，$A_o \in \mathbb{R}^{m\times n_o}$ 为遮挡字典[13,25]，具体可以为单位矩阵、傅利叶基、Haar 小波基等标准正交基[6]，$e \in \mathbb{R}^{n_o}$ 为 y 中的遮挡成分相对于遮挡字典 A_o 的编码. 当 A_o 为单位阵 $\boldsymbol{I} \in \mathbb{R}^{m\times m}$ 时，Wright 等人[23]进一步证明了 RSRC 的鲁棒性：即使待识别图像 y 的像素点近于 100%受到噪声干扰，只要 y 的维数 m 足够高，精确恢复重构系数 x 仍然是可能的，并且在计算上是可行的. 然而，实际中采集到的人脸图像往往是低维的. 在低维情形下，Wright 等人[6]的实验表明不同的特征对 RSRC 的分类能力还是有显著差别的. 因此，基于 RSRC 模型进行识别时往往仍需考虑特征提取和特征选择的问题. 另外，高维会带来计算上的困难，尤其是 SRC 和 RSRC 都是建立在最小 ℓ_1 范数的基础上的：由于 ℓ_1 范数不可导，式(2)不存在解析解. 对 RSRC 模型而言，还需要考虑标准正交的遮挡字典 $A_o=\boldsymbol{I}$ 的维数会随着样本维数的增高而呈指数级增长，进而导致求解 RSRC 往往需要高昂的存储开销和计算开销.

基于计算性能的考虑，Zhang 等人[24]重新审视 SRC 模型(1)，认为 SRC 模型的有效性并非一定是建立在稀疏编码的基础上，提出了“协同表示”的思想：由于不同类别的人脸图像之间存在较大的相关性，不能够被本类别的样本充分表示的局部特征可能能够被其它类别的具有相似局部特征的样本协同表示(如图 2(a)所示的测试样本 y 的“张嘴”表情的例子). 然而，仅仅是协同表示只能保证测试样本 y 被整个训练字典 A 充分表示，并不能保证其重构系数 x 的主要能量一定能落在 y 所属的子空间中(如图 2(b)所示的测试样本 y 的“围巾”遮挡的例子). 为了提升同类样本对测试样本 y 的表达力度，必须“抑制”重构系数 x 在其它类别的样本上因为局部特征的相似而造成的“过度表达”. Zhang 等人[24]提出的基于协同表示的分类(Collaborative Representation-based Classifier，CRC)模型如下：

$$\min_{x,e}\|e\|_2+\lambda\|x\|_p,\ \text{s.t. } y=[AA_o]\begin{bmatrix}x\\e\end{bmatrix} \quad (3)$$

其中，对重构系数 x 的约束采用 ℓ_p ($p=1,2$)范数，λ 为“抑制”系数，通常取较小的值. Zhang 等人[24]的实验表明，CRC 的分类性能足以与 RSRC 相媲美，但 CRC 由于采用了最小 ℓ_2 范数回归，计算复杂度大大降低了. 图 2 比较了 RSRC 与 CRC 的协同表示能力与抗“混叠”的能力. 有趣的是，测试样本中的有些局部特征(如“张嘴”的表情)能够被 RSRC 和 CRC 有效协同表示并抑制，而有些局部特征(如“围巾”遮挡)反而被 RSRC 和 CRC 放大了. 那么，哪些局部特征能够被有效协同并抑制，哪些不能够？这是尚待解决的问题.

2.2　遮挡字典的学习与压缩

在子空间回归模型(2)和(3)中，除了需要考虑对回归系数的约束外，还需要考虑遮挡字典 A_o 的构造和设计. 除了 Wright 等人[6]建议的单位阵、傅利叶基、Haar 小波基等标准正交基外，已有的遮挡字典主要有：有监督的遮挡字典[26]、类间扰动字典[20]、基于投影误差学习的遮挡字典[22]、Gabor 遮挡字典(Gabor Occlusion Dictionary，GOD)[13,25]等，其中，类间扰动字典[20]和基于投影误差学习的遮挡字典[22]的构造都需要已知测试样本中的遮挡信息，本文暂不予讨论.

由于遮挡的颜色和位置等信息是先验未知的，RSRC 中设置遮挡字典 A_o 为通用的标准正交基. 然而，由于人脸图像的维数通常较高，这会导致标准正交的遮挡字典变得非常庞大. 另一方面，遮挡通常只占据了人脸图像的一小部分(具有空间局部性和连续性)，用庞大的标准正交的遮挡字典来表示这一小部分内容也是不经济的. 这就需要利用遮挡的空间局部性和连续性对标准正交的遮挡字典进行压缩.

为了充分利用遮挡的空间连续性与局部性，Wei 等人[26]提出了有监督的遮挡字典：令 $A_e=\boldsymbol{I}$，将空间上连续的字典原子赋予相同的类标(如图 3(a)所示)，然后通过对重构系数和重构误差分别施加最小化 $\ell_{2,1}$ 范数约束，从而达到同时对重构系数和重构误差进行结构化聚类的目的. 实际上，Wei 等人[26]所提出的有监督遮挡字典相当于通过对单位字典中的原子局部重组后所形成的一种压缩了的具有块状结构的字典，如图 3(b)所示. 由于所形成的遮挡字典的每个原子所包含的形状块的大小是固定的，基于这种原子结构的遮挡字典不利于描述人脸图像中的具有更小尺度的遮挡信息. Yang 等人[13,25]提出了基于 Gabor 特征的鲁棒表示与分类方法(Gabor feature-based Robust Representationand Classification，GRRC)，其中最关键的是对 Gabor 遮挡字典的处理：利用 Gabor 变换将遮挡字典 $\boldsymbol{I}$ 中的各原子变换到 5 个尺度 8 个方向的 Gabor 域中，得到 Gabor 遮挡字典. 然而，这样形成的 Gabor 字典的维数是单位字典 $\boldsymbol{I}$ 的 40 倍，Yang 等人[13,25]建议利用一致性下采样和 K-SVD 字典学习[41]从冗余的 Gabor 遮挡字典中获取遮挡字典的紧致表示，从而实现对原始遮挡字典的压缩，如图 3(d)所示. 利用固定的遮挡字典来表示现实中各种类型的遮挡仍然

是困难的，本文的实验(详见第 7 节)表明 GRRC 的识别性能在许多情况下弱于其它鲁棒分类方法.

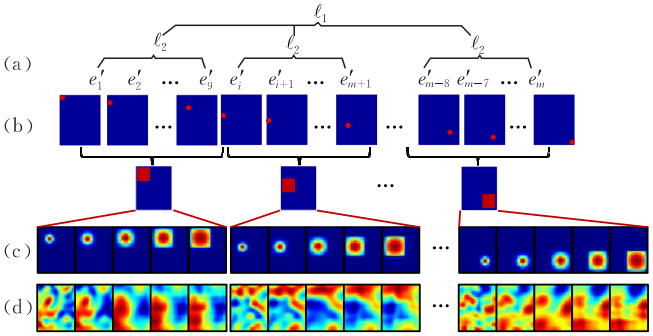


图 3 遮挡字典的压缩过程((a)压缩之前的字典原子为单位向量;(b)对空间中临近的字典原子进行重组所形成的具有块状结构的遮挡字典，等同于 Wei 等人[26]所提出的基于 $\ell_{2,1}$ 约束的有监督遮挡字典;(c)对单位矩阵中的原子进行 Gabor 变换所形成的 Gabor 遮挡字典;(d)对 Gabor 遮挡字典压缩之后所形成的新的遮挡字典，其各原子成分具有更为丰富的形状和方向特征)

3 结构化误差编码

与常见的由高斯噪声所引发的误差不同，由遮挡所引发的误差通常都具有一定的空间结构，因此，对遮挡误差的编码通常都需要考虑遮挡的结构，属于结构化误差编码的范畴. 现有的对遮挡或遮挡误差的结构化编码方法主要有 3 种:(1)通过构造与遮挡结构相匹配的遮挡字典来表示遮挡，也就是 2.2 节所描述的遮挡字典学习方法;(2)另一种与构造遮挡字典相反的方法是通过描述人脸图像本身所固有的结构来表示人脸图像，从而也就分离出了遮挡，此类方法通常与鲁棒子空间学习[42-44]紧密相关，详见 3.1 节;(3)将遮挡自身所固有的结构嵌入重构误差中，直接对结构化误差进行编码[11,27,29]，详见 3.2 节.

3.1 基于鲁棒 PCA 的结构化误差编码

人脸图像之间具有天然的强相关性，这种强相关性可以用“低秩”结构来描述. 而噪声或遮挡的存在会在某种程度上破坏这种低秩结构，借助于人脸图像的低秩结构可以消除噪声. Wei 等人[39]基于鲁棒主成分分析(Principle Component Analysis, PCA)[42]给出了人脸图像训练字典的学习方法. 那么，如何将现有的鲁棒 PCA 方法用于处理“测试集中包含噪声(遮挡)”的识别问题呢? Luan 等人[21]给出了解决此类问题的一个范例:将有遮挡的待识别图像 y 依次与各人脸子空间的训练样本 A^k 联合，形成一个扩展的数据集 $[A^k, y]$，再在 $[A^k, y]$ 上实施鲁棒 PCA，就可以得到有遮挡图像 y 相对于子空间 A^k 的重构图像 $\hat{y}^k$ 和误差图像 e^k. 这里，利用鲁棒 PCA 对有遮挡图像 y 的处理利用了来自于同一个子空间的人脸图像的低秩结构和噪声的稀疏分布. 为了保有“低秩”，必须尽可能地“稀释”掉数据集 $[A^k, y]$ 中的噪声，并使得数据集中的各样本趋同. 这一方面会导致 y 中的遮挡信息向 A^k 的各图像中扩散;另一方面，由于 A^k 中的图像数量占了主导地位，所恢复的重构图像 $\hat{y}^k$ 将与 A^k 中的图像非常相似. 因此，y 的判别信息将主要集中在其误差图像 e^k 中. Luan 等人[21]假设如果 y 与 A^k 中的图像为同一类别，则误差图像 e^k 应该具有更好的稀疏性和光滑性，给出了对 e^k 的稀疏性和光滑性描述，进而给出了 y 到 A^k 的距离度量. 实验表明[21]，该距离度量相对于直接计算误差图像 e^k 的 ℓ_1 范数或 ℓ_2 范数更为有效.

基于鲁棒 PCA 的最近子空间分类方法的识别性能依然依赖于各人脸子空间所能提供的训练样本的数量，如果数量不够将不能很好地“稀释”来自于待识别图像的遮挡信息，同样会导致局部混叠问题. 另一方面，经典的鲁棒子空间学习方法[42,45]都需要首先将二维图像拉伸为列向量，这样会造成图像的空间信息的丢失. 最近，Lu 等人[46]提出了基于张量的鲁棒 PCA 方法，我们期待这一方法能够解决上述问题.

3.2 重构误差中的图像空间信息嵌入

遮挡的空间结构信息对于准确表示因遮挡而引发的误差非常重要. 2.2 节所讨论的字典学习方法[13,25]将这种空间结构信息嵌入在了遮挡字典中;而在 4.2 节将要介绍的 SEC-MRF[12]和 SSEC[9]等模型都对遮挡的空间结构(即遮挡支撑 s)进行了显式的编码. 那么，是否可以直接将遮挡的空间结构信息编码在重构误差 e 中呢?

受结构化稀疏编码[47]思想的启发，Jia 等人[27]提出用基于树形结构的 $\ell_{\infty,1}$ 范数替换 RSRC 模型(2)中对重构误差 e 的 ℓ_1 范数约束:首先，将 e 按照树形结构划分为不同的块;然后，对每一块施加 ℓ_∞ 范数(相当于局部最大的池化操作①);最后，再对所有块的 ℓ_∞ 范数施加 ℓ_1 范数. 与 Wei 等人[26]所提出的对重构误差 e 的 $\ell_{2,1}$ 范数约束(2.2 节)不同的是:

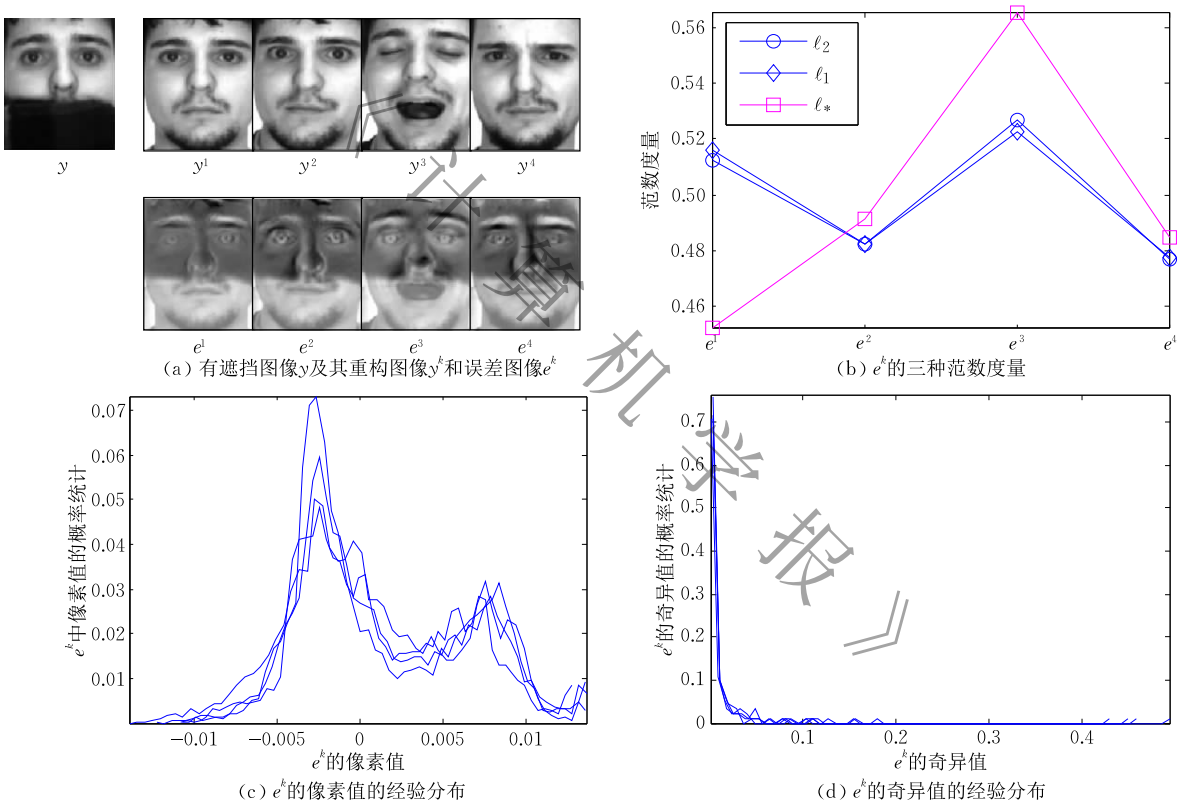
① 池化(pooling)是卷积神经网络(不局限于此)中常用到的一种特征处理技术，用于将空间上临近的或语义上相似的多个特征点合并为一个特征点[78]，其主要作用是产生具有平移不变性的特征和防止因为冗余特征所导致的过拟合问题. 常用到的池化操作有:最大化池化(max pooling)和平均池化(average pooling).

Jia 等人[27]所提出的基于树形结构的 $\ell_{\infty,1}$ 范数约束实际上是对重构误差 e 的多尺度聚类；在 $\ell_{\infty,1}$ 范数中嵌入的 ℓ_∞ 范数是一种非线性池化操作，因此，对重构误差 e 的 $\ell_{\infty,1}$ 范数约束不能转变成对遮挡字典的约束.

Qian 等人[29]提出另一种将图像的空间结构信息嵌入到重构误差 e 的编码过程中的方法：基于核范数（ℓ_* 范数，即矩阵的奇异值之和）正则化的回归方法（Robust Nuclear norm regularized Regression，RNR）. RNR 对重构误差 e 同时施加 ℓ_2 范数和 ℓ_* 范数两种约束：

$$\min_{x,e}\|e\|_2+\gamma\|\mathcal{M}(e)\|_*+\lambda\|x\|_2,\ \text{s.t.}\ y=Ax+e \quad (4)$$

其中，$\gamma\geq 0$ 为正则化参数，$\mathcal{M}(e)$ 表示将图像 e 的列向量表示还原为矩阵表示. ℓ_* 范数与常规的 ℓ_p 范数相比有两点优势：(1) ℓ_* 范数能够更好地“感知”图像的局部形变，如图 4(a)～(b)所示；(2) ℓ_* 范数具有更好的抗噪声干扰的能力. ℓ_* 范数逼近于矩阵的秩[42-43]，度量的是图像在其主方向上的能量（奇异值），而主方向上的能量通常是稀疏的且去除了一定的噪声干扰. 图 4(c)和(d)比较了图 4(a)中的四幅重构误差图像的经验分布及其奇异值分布，可以看出，奇异值分布更有规律也更加稀疏. 事实上，Kim 等人[48]于 2104 年提出的 SVD 脸也正是利用奇异值分解来去除图像中的噪声影响的.



(a) 有遮挡图像y及其重构图像y^k和误差图像e^k　(b) e^k的三种范数度量

(c) e^k的像素值的经验分布　(d) e^k的奇异值的经验分布

图 4　误差的经验分布与度量

差不多在与 Qian 等人[29]的工作相同的时间内，Luo 等人[11]提出了类似的方法. 所不同的是，Luo 等人[11]认为使用单一的高斯分布或拉普拉斯分布很难拟合实际的误差分布，提出进一步将重构误差 e 分解为稀疏无界的误差 e° 和有界但不一定是稀疏的误差 $e^\cdot$，并对 $e^\cdot$ 施加 ℓ_2 范数约束，对 e° 同时施加 ℓ_* 和 ℓ_p 范数（$p=1$ 或 2）约束.

基于分块的树形结构的 $\ell_{\infty,1}$ 范数[27]和混合的 ℓ_p/ℓ_* 范数[11,29]都嵌入了重构误差 e 的空间结构信息，尽管二者并没有对遮挡进行显式处理，但将它们与子空间回归方法相结合都表现出了良好的识别性能. 这意味着遮挡信息在正确规则的引导下可以被隐式地编码并消除，5.3 节将进一步印证这一观点.

4　迭代重权误差编码

子空间回归方法和结构化误差编码方法都试图用已有的字典或遮挡的结构来表示人脸图像所包含的遮挡，是一种典型的“生成模型”. 然而，视觉经验告诉我们[9]：人眼在识别一幅有遮挡的人脸图像时通常只是简单地忽略了遮挡区域的内容，根据未遮挡区域进行识别. 因此，人眼只是在需要识别的时候

才调用“生成模型”，对遮挡的处理则是一种典型的“判别模型”：只简单地区分是或不是遮挡，而不去理解遮挡的内容.加权的误差编码模型[9-10,12,14,19]恰类似于人眼的识别机制：通过学习权重向量 $\boldsymbol{\omega}\in[0,1]^m$ 或遮挡支撑 $s\in\{-1,1\}^m$ 来抑制或消除遮挡，然后再通过对未遮挡区域编码来实现分类.由于一般情况下，求解加权的误差编码模型需要通过权重和误差的交替迭代优化，因此，我们将此类模型统称为“迭代重权误差编码”，具体如 4.1 节的式(8)和式(9)所示.

4.1 基于鲁棒误差度量和鲁棒回归的迭代重权误差编码

2011 年，由 He 等人[10]提出的基于相关熵[49]的稀疏表示(CorrEntropy-based Sparse Representation，CESR)模型和 Yang 等人[14]提出的鲁棒稀疏编码(Robust Sparse Coding，RSC)是两个最具有代表性的加权误差编码模型，目前已被作为一种基准方法，在鲁棒人脸识别领域被广泛使用和比较[9,11,19,29,32].CESR 和 RSC 分别从鲁棒误差度量与鲁棒回归的角度描述了重构误差 e 与权重 $\boldsymbol{\omega}$ 之间的关系.由于鲁棒误差度量与鲁棒回归的天然联系，两者在本质上是等价的，可统一描述如下：

$$\min_{x,e,\boldsymbol{\omega}}\|\boldsymbol{\omega}\odot e\|_2^2+\lambda\|x\|_p+\gamma\cdot\varphi(e,\boldsymbol{\omega})\quad \text{s.t. } e=y-Ax \tag{5}$$

其中，$\odot$ 表示 Hadamard 积(即两个向量中对应位置上的元素的乘积)，$\|\cdot\|_p$ 表示 $\ell_p(0<p\leq 2)$ 范数，$\lambda\geq 0$ 和 $\gamma\geq 0$ 为正则化参数；$\boldsymbol{\omega}\in[0,1]^m$ 为权重向量，用以抑制 y 中的坏特征；$\varphi(e,\omega)$ 表示通过重构误差 e 来估计权重向量 $\boldsymbol{\omega}$ 的代价函数.

子空间回归模型(1)～(3)对重构误差 e 的度量主要是基于 ℓ_p 范数($p=1,2$)的.基于 ℓ_p 范数的误差度量的主要问题在于将两幅图像的所有像素点的差异等同对待，而人眼在辨别两幅图像是否相似时主要是根据它们相似的区域，而非不相似的区域[50-51].因此，对于有遮挡图像而言，基于局部相似性的误差度量十分关键，这正是加权误差编码模型(5)所强调的.基于相关熵的度量(Correntropy Induced Metric，CIM)[49]可以自动做到局部度量：

$$\mathrm{CIM}(e)\triangleq 1-\frac{1}{m}\sum_{i=1}^{m}g(e_i) \tag{6}$$

其中，$g(e_i)=\exp(-e_i^2/(2\sigma^2))$ 为高斯核函数.CIM 自动将重构误差 e 划分为三个大小不同的区域：欧几里德区域、过渡区域和校正区域.这一区域的划分相当于对不同大小的 e_i 施加了不同的权值：较小的误差被赋予了较大的权值，而较大的误差被赋予了较小的权值.基于 CIM，He 等人[10]提出了 CESR 模型：

$$\min_{x,e}\mathrm{CIM}(e)+\lambda\|x\|_1,\ \text{s.t. } e=y-Ax,x\geq 0 \tag{7}$$

由于误差权重 $\boldsymbol{\omega}$ 是隐含在 CIM 中的，CESR 没有明确定义式(5)中的 $\varphi(e,\boldsymbol{\omega})$.CESR 在形式上与鲁棒回归模型[52-53]是等价的，而迭代重权最小二乘是求解鲁棒回归模型的一种有效方法，具体迭代过程如下：

$$\boldsymbol{\omega}_i^{(t+1)}=g(e_i^{(t)}) \tag{8}$$

$$(x^{(t+1)},e^{(t+1)})=\arg\min_{x,e}\|\omega^{(t+1)}\odot e\|_2^2+\lambda\|x\|_1,\quad \text{s.t. } e=y-Ax,\ x\geq 0 \tag{9}$$

其中，式(8)对权重 $\boldsymbol{\omega}_i$ 的估计基于凸共轭函数理论[54].我们把形如式(8)和式(9)的迭代模型称为“迭代重权误差编码”.

差不多在与 CESR[10]提出的相同的时间内，Yang 等人[14]从误差概率分布的角度出发也提出了类似于式(8)和式(9)的迭代重权误差编码模型.Yang 等人[14]指出现实误差的分布可能远非某种特定的分布所能描述，从鲁棒回归的角度出发，给出了误差分布的概率密度函数应该遵循的三个假设：一阶可导、关于原点对称、误差绝对值越大则概率密度值越小.基于这些假设，Yang 等人[14]采用了如下基于 Logistic 函数的概率密度函数

$$p_{\mathrm{RSC}}(e_i\mid\mu,\delta)=(1+\exp(\mu(e_i^2-\delta)))^{-1} \tag{10}$$

其中，δ 和 μ 为控制参数：δ 类似于高斯分布中的参数 σ，控制着“钟形窗口”的大小；μ 控制着 $p_{\mathrm{RSC}}(e_i)$ 的值从 1 下降到 0 的速度，当 μ 较大时，p_{RSC} 为矩形窗.图 5 比较了不同参数下的 $g(\cdot)$ 和 $p_{\mathrm{RSC}}(\cdot)$，由于 $p_{\mathrm{RSC}}(\cdot)$ 有两个控制参数，它比 $g(\cdot)$ 更为灵活.基于最大后验估计和一阶泰勒展开式，Yang 等人[14]得

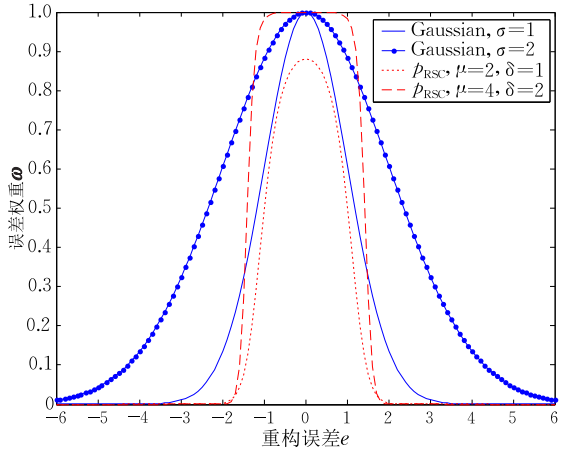


图 5 CESR 和 RSC 中误差权重 $\boldsymbol{\omega}$ 与重构误差 e 的关系：高斯核函数 $g(\cdot)$ 与 $p_{\mathrm{RSC}}(\cdot)$ 函数的比较

到了类似于式(8)和式(9)的迭代重权误差编码模型：RSC 模型. 与 CESR 最主要的不同是，RSC 采用了基于 p_{RSC} 的误差权重估计：

$$\boldsymbol{\omega}_i = p_{\mathrm{RSC}}(e_i) \tag{11}$$

CESR 和 RSC 虽然都采用了形如式(8)和(9)的迭代重权方法求解目标式，但它们求解重构系数 x 和误差权重 $\boldsymbol{\omega}$ 的优化策略不同. CESR 将求解权重 $\boldsymbol{\omega}$ 的过程嵌入了求解重构系数 x 的过程中：一边通过已经求得的部分 x 更新 $\boldsymbol{\omega}$，一边将 $\boldsymbol{\omega}$ 用于噪声抑制来进一步校正 x 的求解结果，直至收敛；而 RSC 则是将 x 的求解过程看成了黑盒子，根据式(9)求得 x 的一个迭代解之后，再根据式(11)更新 $\boldsymbol{\omega}$，交替迭代，直至收敛. 显然，CESR 的计算速度更快，但哪种方法的识别性能更好呢？本文第 7 节的实验表明：CESR 的识别性能在鲁棒特征域中往往优于 RSC，而在原始像素域中这种性能优势并不能保证. 这是因为在原始像素域中因噪声而引发的局部混叠容易导致 CESR 在一开始迭代时因为没有获得全局重构信息而错误地“强调”噪声特征，并且在后续迭代过程中，这种错误可能被继续“强调”，得不到纠正；而在鲁棒特征域中，由于噪声特征已经在一定程度上被预处理了，CESR 的加权机制又使其可以及早抑制仍然残留的噪声特征，从而达到更好的识别效果. 对于 RSC 而言，不管是在像素域还是在鲁棒特征域中，由于每次迭代都需要等到完整求得回归系数之后才会通过权重抑制噪声，所计算的回归系数可能因为噪声的干扰在一开始迭代时已经偏离真实值较远了.

另外，CESR 和 RSC 主要存在如下三个方面的问题：(1) 它们都是通过权重来抑制坏特征，而权重主要是根据误差来计算的：误差越大权重越小，但没有提供任何有力的技术来保证遮挡所造成的误差一定是较大的，本文实验部分的图 21 和图 22 对此进行了分析；(2) 尽管 CESR 和 RSC 都采用了与人类视觉相似的局部误差度量，但与人类视觉不同的是，它们所采用的局部误差度量，没有充分考虑到人类视觉的 Weber 效应[34]，即：对于相同的误差施加了相同的权重，没有考虑到在进行误差比对之前的像素值的大小(本文实验指出将鲁棒特征嵌入 CESR 和 RSC 可以获得极大的性能提升)；(3) CESR 和 RSC 所采用的局部误差度量只具有统计意义上的局部性，而非空间意义上的，完全丢掉图像的空间信息可能会引发在视觉上并不存在的混叠，我们在 3.2 节讨论了空间信息的误差嵌入，在 4.2 节将进一步讨论与图像的空间结构相关的误差编码方法.

关于 CESR 和 RSC 的后续工作主要有：2014 年，He 等人[31]将 CESR 模型(7)拓展为一般的基于 M-估计子的鲁棒误差编码框架，并在该框架下研究误差校正方法和误差检测方法的关系；2013 年，Yang 等人[16]基于协同表示的思想对 RSC 进行了扩展，提出了正则化鲁棒编码方法(RRC)，与此同时，他们又将鲁棒特征嵌入 RSC，提出了基于鲁棒核表示[32]的分类方法(详见第 6 节).

4.2　基于 Markov 随机场的遮挡支撑估计

解决遮挡问题最直接的方法是检测到遮挡的位置并将其从进一步识别中排除. 要准确定位遮挡，在直觉上比较合理的方法是为遮挡与正常人脸的“差异”充分建模. 然而，近年来的研究表明，除了考虑遮挡与人脸的差异外，还需要充分考虑遮挡自身所具有的先验信息. 目前所知道的现实遮挡的具有普适性的先验信息主要是遮挡的空间局部性和连续性[12,55-56]以及遮挡边缘的规则性[9]，将这些共性与误差编码模型相结合产生了一系列有遮挡人脸图像的鲁棒修复与识别算法[9,12,19,55-56]，其中，以 Zhou 等人[12]于 2009 年提出的基于 Markov 随机场(Markov Random Field，MRF)的稀疏误差校正模型(Sparse Error Correction with MRF，SEC-MRF)和 Li 等人[9]于 2013 年提出的结构化稀疏误差编码(Structured Sparse Error Coding，SSEC)模型最具有代表性.

MRF 是目前唯一的被广泛用于描述遮挡的空间局部连续性的模型[9,12,19,55-56]. 遮挡的 Markov 性主要体现在：当前像素点是否为遮挡点只跟其邻域像素点的状态相关，而与距离较远的像素点的状态无关. 通常用遮挡支撑 $s\in\{-1,1\}^m$ 来描述有遮挡人脸图像 y 中的各像素点的状态：$s_i=-1$ 表示 y_i 为“非遮挡”状态，$s_i=1$ 表示 y_i 为“遮挡”状态. 在二维平面内，遮挡的 Markov 性可以用 MRF 模型来描述，如图 6(a)所示，是一个带权的无向图，可用 Ising 模型描述如下：

$$p(s)\propto\exp\Big(\sum_{i=1}^{m}\lambda_i s_i+\sum_{i=1}^{m}\sum_{j\in\mathcal{N}(i)}\lambda_{ij}s_i s_j\Big) \tag{12}$$

其中，$\mathcal{N}(i)$ 表示 s_i 的邻域节点的索引的集合，λ_i 为数据费用参数，用来衡量将 s_i 设置为某种状态(−1 或 1)的代价，λ_{ij} 为平滑费用参数，用来衡量将 s_i 的状态迁移到 s_j 的状态的代价. Ising 模型(12)实际上是遮挡支撑 s 的概率生成模型，通过最大化该生成模型就可以获取在给定参数(λ_i 和 λ_{ij})下的遮挡支撑 s 的最优解.

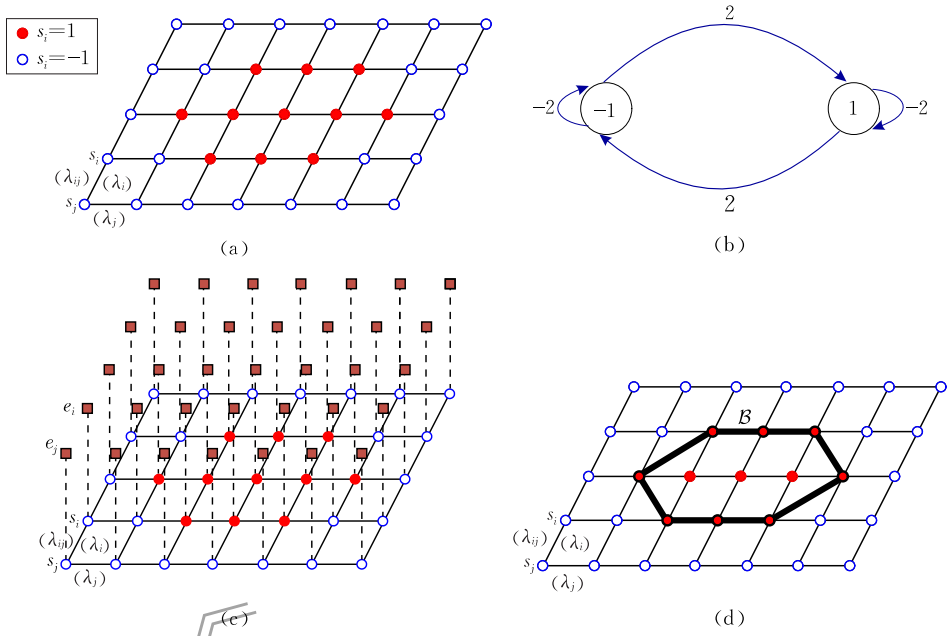


图 6 遮挡支撑 s 的图模型((a) MRF;(b) MRF 中各节点的状态迁移图,箭头上方的数字为状态迁移的费用;(c)从重构误差到 MRF 的映射,用以建立重构误差相对于遮挡支撑的似然函数;(d)形态图,一个带权的无向图,用以描述遮挡边缘的规则性,由顶点、边和子图边缘三个部分组成,图中黑色线条所连接的顶点为 $s_i=1$ 的子图的边缘)

显然,Ising 模型中的参数 λ_i 和 λ_{ij} 将决定着所生成的遮挡支撑 s 的质量. 一般来说,λ_{ij} 容易根据经验来确定,只需要给出各种状态彼此迁移的费用即可,如图 6(b)所示. 而 λ_i 的确定则需要十分谨慎,它是连接有遮挡图像 y 和遮挡支撑 s 的桥梁,通常需要根据 y 及其重构 $\hat{y}$ 的似然函数 $p(y_i|\hat{y}_i,\theta)$ 来确定. Zhou 等人[12]基于稀疏误差校正重构图像,其实质为形如式(5)的加权误差编码模型:

$$\min_{x,e,s}\|s'\odot e\|_1+\lambda\|x\|_1-\gamma\log p(e,s),\ \text{s.t.}\ e=y-Ax \tag{13}$$

其中,$s'=(1-s)/2$,$p(e,s)$为重构误差 e 和遮挡支撑 s 的联合概率密度函数. 通过 $p(e,s)=p(e|s)p(s)$ 可以将误差编码模型(13)和 Ising 模型(12)连接起来,而通过似然函数 $p(e|s)$可以将 e 和 s 连接起来. 由于 e 和 s 的所有元素存在一一映射关系,如图 6(c)所示,可以通过 e 和 s 的似然函数 $p(e|s)$确定 Ising 模型中的数据费用参数 λ_i. Zhou 等人[12]所建立的 $p(e|s)$可简单理解为从重构误差 e 到遮挡支撑 s 的映射函数:$s_i^0=\mathcal{K}_\tau(e_i)$,其中,$\mathcal{K}_\tau(e_i)=\begin{cases}-1, & |e_i|\le\tau\\ 1, & |e_i|>\tau\end{cases}$为硬阈值滤波,$\tau$ 为给定的经验阈值. 根据 s_i^0就可以设置参数λ_i:如果 $s_i^0=1$,可以令 λ_i 为一个较大的值(如 $\lambda_i=2$);如果 $s_i^0=-1$,可以令 λ_i 为一个较小的值(如 $\lambda_i=-2$).

在确定了参数 λ_i和λ_{ij}之后,就可以通过最大化 Ising 模型(12)进一步估计遮挡支撑 s,然后,再由式(13)重新估计重构误差 e. e 和 s 交替迭代,最终获取它们的稳定值.

在 SEC-MRF 中,阈值 τ 的选取对迭代结果会有很大的影响. Zhou 等人[12]建议 τ 可以取一个较大的初值,然后在迭代过程中逐渐收缩 τ 的值. 这一建议是假设随着迭代次数的增加,有遮挡图像 y 的重构也更加精确了. 然而,这种假设并不一定成立,最优的重构并不一定是由最后的几次迭代产生的,如图 7 的第 2 行所示.

为了校正 SEC-MRF 的迭代结果,Li 等人[9]基于形态图模型提出了 SSEC 模型. 形态图模型(如图 6(d))与传统的图模型(如图 6(a))相比多了一个子图边界描述,用以对遮挡的边界建模. Li 等人[9]注意到现实世界的遮挡往往具有非常规则的边缘轮廓. Ising 模型(12)是将 MRF 映射到了传统的图模型上,Li 等人[9]将 MRF 映射到了形态图模型上,得到了新的遮挡支撑的生成模型:

$$p(s)\propto\exp\Big(\sum_{i=1}^{m}\lambda_i s_i+\sum_{i=1}^{m}\sum_{j\in\mathcal{N}(i)}\lambda_{ij}s_i s_j-\lambda_{\mathcal{B}}\sum_{i\in\mathcal{B}}s_i\Big) \tag{14}$$

其中,$\lambda_{\mathcal{B}}\ge 0$ 为遮挡边缘正则化参数,$\mathcal{B}$为位于遮挡子图边界上的顶点的索引的集合. 将式(14)代入式(13)就得到了 SSEC 模型. 从模型上来看,SSEC 与 SEC-MRF 的主要区别在于两者的遮挡支撑 s 的生成模型不同,但这一区别促使 SSEC 在优化过程中采取了更加灵活的遮挡支撑估计评估方案:

$$\min_t c_E^{(t)}+\lambda_B c_B^{(t)} \tag{15}$$

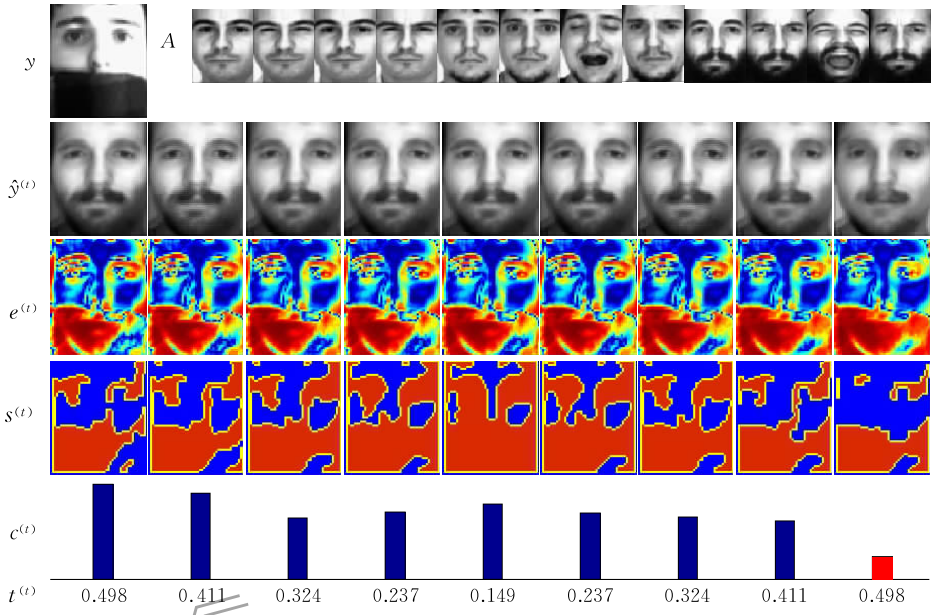


图 7　SSEC 模型[9]的人脸重构与遮挡检测过程($c^{(t)}=c_E^{(t)}+\lambda_B c_B^{(t)}$ 为 SEEC 的遮挡支撑质量评估准则)

其中，$c_E^{(t)}$ 和 $c_B^{(t)}$ 分别为第 t 次迭代所产生的非遮挡区域的重构误差的能量和遮挡边界的长度. 灵活的遮挡支撑评估方案也促使 SSEC 采用了更加灵活的阈值选取策略，如图 7 的最后一行所示. 有趣的是，这一策略往往导致重构图像的质量在阈值 $\tau^{(t)}$ 回升的过程中得到提升. 图 7 中的柱状图展示了 SSEC 如何根据式(15)度量并选取最优的遮挡支撑.

显然，在遮挡检测的过程中，阈值 τ 的选取十分重要. 从图 7 中显示的遮挡检测结果来看，尽管 SSEC 采取了非常谨慎的阈值选取策略，其最优的遮挡支撑估计在眉毛和眼睛周围仍然出现了误检，而在有围巾遮挡的区域则出现了漏检. 这是因为 SSEC 与 SEC-MRF 一样在每次迭代时对所有像素点上的重构误差都采用了相同的硬阈值进行滤波. 显然，这样做是不合理的，Yang 等人[57]近来提出了基于字典学习的残差图学习方法，用以自动获取与图像各个部位的像素值相适应的阈值.

5　鲁棒特征提取

越来越多的特征提取[13,19,25-26,32-34,58-63]和学习方法[17-18,64-65]正在应用于有遮挡人脸识别并极大地提升了识别性能. 那么，为什么这些特征是对遮挡鲁棒的？鲁棒的特征提取应该遵循什么原则？浅层特征和深度特征有什么不同？从浅层特征到深度特征经历了怎样的变化？本节将通过对现有方法的回顾来探讨上述问题.

5.1　鲁棒特征提取的基本要素

一幅人脸图像所包含的特征通常极为丰富，既包括颜色、亮度、纹理、方向等低阶特征，也包括姿态、表情、年龄、人种等高阶特征. 鲁棒的特征提取方法通常需要对这些特征进行分解，例如，Gabor 变换[66-69]将图像分解到多个尺度和多个方向上，而属性学习[70-73]将图像分解为多个可描述的属性. 那么，为什么需要对图像进行分解呢？首先，图像的各种特征以非线性的方式组合在一起，容易相互干扰，特别是当噪声特征与正常特征混合在一起时，噪声特征可能会对分类结果起到主导作用(如图 2 所示的“局部混叠”现象)；其次，机器视觉与人类视觉[74]相似，都需要对待识别的图像进行“表示”(由什么成分组成)，然后才知道它到底是“什么”，而图像表示是一个极其复杂的从单个像素到局部结构、从低阶语义到高阶语义的分解过程.

视觉神经科学的研究表明人类视觉皮层“认知”视觉信号的过程大致经历了 LGN-V1-V2-V4-IT 等 5 个阶段[75]，每一个阶段都在前一个阶段的基础上对输入信号的特征进行编码和组合，形成更高阶的特征. 从人类视觉系统对图像处理的最终结果来看，高阶语义特征比低阶语义特征更为自然，也更具有不变性[18,70].

显然，并非越高阶的语义特征越有利于识别，例如，全局性的特征提取方法，如主成分分析(PCA)[76]、非负矩阵分解(NMF)[77]等，都是以整幅图像为分解单元的. 这些全局特征不如局部特征鲁棒且不适

用于有遮挡人脸识别问题[6]. 可见,为了避免局部误差的扩散,仅仅对图像进行高阶语义分解还不够,还需要对这种分解施加空间局部性约束. LeCun 等人[78]指出图像的空间局部特征高度相关能够形成独特的主题. 事实上,对传统的特征提取方法施加空间局部约束,往往会取得令人惊讶的效果,例如,局部非负矩阵分解[79]和 PCANet[17]所使用的滤波器组正是分别对已有的方法(即 NMF 和 PCA)施加了空间局部约束后得到的,它们相对于原有方法都取得了显著的性能提升.

综上,可以认为图像分解、高阶语义特征提取、空间局部性与多层次处理是鲁棒特征提取的四个基本要素.

5.2 从光照鲁棒到遮挡鲁棒

由于光照变化也可以被视作一种笼罩在人脸图像表面的"半透明"的遮挡,因此,对光照鲁棒的特征也常常被用于处理遮挡问题. 目前,受到广泛关注的具有光照不变性的遮挡鲁棒特征主要有:图像梯度方向(Image Gradient Orientation, IGO)特征[33,80]和 Weber 脸[34,81]. IGO 特征[33],又称为梯度脸[80],在人脸识别领域中,首先被用于处理有光照变化的人脸图像[80],此后,受到了广泛的关注[15,33-34,60,62-63],其中一个主要的应用是将量化后的 IGO 应用于图像分解[34,62-63]. Weber 脸(Weberfaces)由 Wang 等人[81]于 2011 年提出,其实质为 Chen 等人[34]所提出的韦伯局部描述子中的图像差分激励. 由于 Weber脸所采用的拉普拉斯滤波器为二阶差分算子,能够更好的刻画图像的边缘和纹理信息,一般被认为具有比 IGO 更好的光照鲁棒性[81].

IGO 和 Weber 脸随后都被应用于处理遮挡问题[7,15,19,26,33,58]. 此外,用于处理光照问题的对数变换[82-83]也被 Li 等人[9]用于设计鲁棒的误差度量算子. 那么,为什么对光照鲁棒的特征通常也对遮挡鲁棒呢?除了光照变化与遮挡在直觉上的联系外,Tzimiropoulos 等人[15,33]关于 IGO 的工作对此做出了部分理论上的解释.

Tzimiropoulos 等人[33]通过引入余弦核将 IGO 拓展为适用性更广的鲁棒特征,尤其是对遮挡问题的处理非常简洁、巧妙、富有哲理. 他们的工作主要基于如下发现:自然界中完全不同的两幅图像 u 和 v 在 IGO 域 $\phi(\cdot)$ 中的差值 $\hat{e}=\phi(u)-\phi(v)$ 以极高的置信度服从 $[-\pi,\pi)$ 上的均匀分布. 由于对称区间上服从均匀分布的各变量的数值期望为 0,因此 $\sum_i \hat{e}_i \simeq 0$. 这一结论揭示了自然界中不同图像之间的"零和差异"现象:两幅不同的图像在单个像素点的差异可能非常大,但总体差异之和为零. 显然,两幅完全相同的图像的差值也是满足"零和差异"的. 为了区分这两种"零和差异",Tzimiropoulos 等人[15,33]在 IGO 中引入了余弦核函数,将两幅图像的相似度定义为

$$\mathcal{S}_\phi(u,v)\triangleq\sum_i \cos(\phi(u_i)-\phi(v_i)) \tag{16}$$

容易证明:

$$\mathcal{S}_\phi(u,v)=\mathcal{G}(u)^{\mathrm{T}}\mathcal{G}(v) \tag{17}$$

其中,$\mathcal{G}(\cdot)=[\cos(\phi(\cdot))^{\mathrm{T}}\ \sin(\phi(\cdot))^{\mathrm{T}}]^{\mathrm{T}}$. 因此,式(16)相当于首先将输入图像分解到了高维特征空间 $\mathcal{G}(\cdot)$ 中,然后再在高维空间中通过内积计算相似度. 显然,两幅完全相同的图像的相似度等于它们的像素维;而两幅完全不同的图像,由"零和差异"假设,其相似度为 0,也就是 $\mathcal{G}(u)^{\mathrm{T}}\mathcal{G}(v)=0$. 如果将完全不同于 u 的 v 视作 u 的遮挡,则有:遮挡图像与人脸图像在 $\mathcal{G}(\cdot)$ 中正交;如果 v 与 u 仅有局部差异,则 $\mathcal{S}_\phi(u,v)=m-\delta$,其中 δ 为 v 和 u 中有差异的像素点的个数. 因此,可以用

$$\rho_\phi(u,v)\triangleq(m-\mathcal{S}_\phi(u,v))/m \tag{18}$$

来估计 v 相对于 u 发生遮挡的比例.

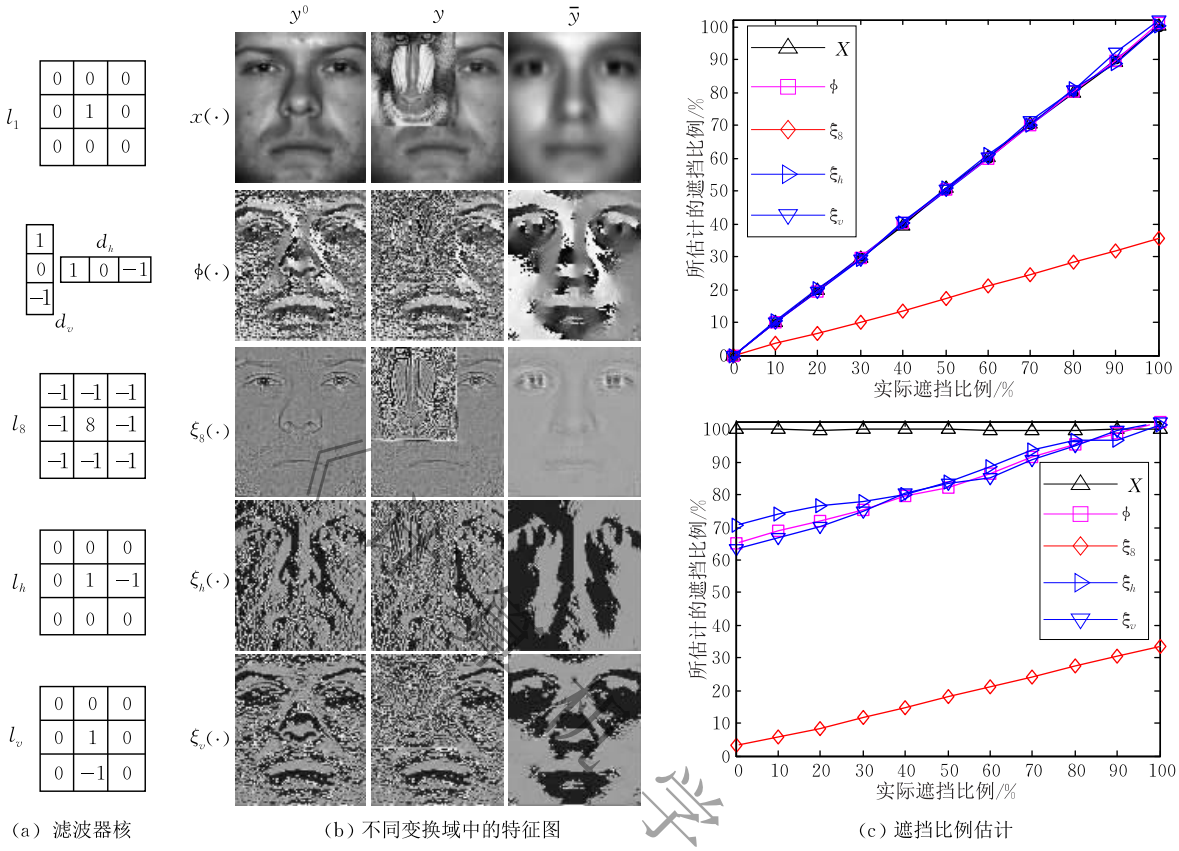
事实上,也可以将式(16)和式(18)拓展为更一般的情形:

$$\mathcal{S}_f(u,v)\triangleq\sum_{i=1}^{m}\cos(f(u_i)-f(v_i)) \tag{19}$$

$$\rho_f(u,v)\triangleq(m-\mathcal{S}_f(u,v))/m \tag{20}$$

其中,$f(\cdot)$ 表示任意的特征变换算子. 显然,并非任意特征变换都会触发"零和差异",但是一个普遍的规律是:只要保证两幅图像的不相同区域在特征域中有足够丰富的差异,就会以较高的概率触发该不相同区域的"零和差异". 这意味着如果 $f(\cdot)$ 为基于卷积滤波的特征变换算子,所采用的滤波器核需足够小. 图 8(a)给出了拉普拉斯滤波器的两个收缩版:l_h 和 l_v,相应的 Weber 特征变换分别用 $\xi_h(\cdot)$ 和 $\xi_v(\cdot)$ 表示. 显然,没有滤波器比单位脉冲滤波器(对应于图 8(a)中的 l_1)更小. 图 8(c)演示了在不同特征域中利用式(20)对具有不同遮挡百分比(从 0~100%)的图像进行遮挡比例的估计. 图 8(c)的上图显示,如果已知有遮挡图像 y 的真实图像 y^0,在 Weber 域中,采用收缩的 l_h 和 l_v 滤波器比直接采用 l_8 滤波器能更好地估计遮挡比例,且与 IGO 域中的估计结果非常接近;直接使用原始的像素特征(对应

于 l_1 滤波）也达到了类似的效果. 但是使用 l_1 滤波得到的是原始像素特征，由于没有考虑到邻域像素点之间的关联，一旦当前像素点与其重构图像对应位置的像素点略有差异，就会认为两者不同. 图 8(c)的下图显示，如果 y 的重构图像为均值脸 $\bar{y}$，那么，不论 y 中是否存在遮挡，在像素域中都始终认为 y 被 100%遮挡了，而在 IGO 域和 Weber 域中对遮挡比例的估计则相对客观.



(a) 滤波器核　　(b) 不同变换域中的特征图　　(c) 遮挡比例估计

图 8　通过对有遮挡人脸图像的遮挡比例的估计来验证不同特征域中的“零和差异”现象（$x(\cdot)$、$\phi(\cdot)$、$\xi_8(\cdot)$、$\xi_h(\cdot)$、$\xi_v(\cdot)$分别表示像素域、IGO 域和三种 Weber 域.(a)不同变换域所用到的滤波器核；(b)有遮挡图像 y 及其真实的无遮挡图像 y^0 和均值脸图像 $\bar{y}$（用以表示一般的人脸结构）在不同特征域中的特征图；(c)在不同特征域中，分别以 y^0（上图）和 $\bar{y}$（下图）为参照，由式(20)估计的 y 中的遮挡比例）

鉴于 $\xi_h(\cdot)$ 和 $\xi_v(\cdot)$ 对遮挡的鲁棒性，与 IGO 中的 $\mathcal{G}(\cdot)$ 相对应，可以定义 Weber 域中的高维特征空间

$$\mathcal{W}(\cdot)=[\xi_h(\cdot)^{\mathrm{T}}\xi_v(\cdot)^{\mathrm{T}}]^{\mathrm{T}} \tag{21}$$

$\mathcal{G}(\cdot)$ 和 $\mathcal{W}(\cdot)$ 实际上都是基于“零和差异”准则对原始图像进行分解的结果，本文的实验表明将 $\mathcal{G}(\cdot)$ 和 $\mathcal{W}(\cdot)$ 嵌入已有的分类器如 CESR 和 RNR 等，能够极大提升其对有遮挡图像的识别性能.

5.3　PCANet 与 FPH 框架

PCANet 由 Chan 等人[17]于 2015 年底提出，其最重要的贡献是将子空间学习引入深度学习，为卷积神经网络(Convolutional Nerual Network，CNN)中卷积核的学习提供了新思路；又将 CNN 的卷积层引入经典的“特征图(Feature Map)-模式图(Pattern Map)-柱状图(Histogram)”的特征提取框架（本文将这一框架简称为“FPH 框架”），将深度学习与传统的特征提取方法建立起了联系. 尽管并没有对光照变化和遮挡做任何先验性假设和显式处理，甚至没有用到大规模的训练数据，PCANet 的神经元响应却对光照变化和遮挡等表现出了很强的鲁棒性. 这一方面再次印证了深度学习强大的异常特征处理能力，另一方面也指出了在深度学习日渐趋于主导地位的今天，传统的基于手工设计的特征提取技术与深度学习之间的联系仍然值得我们思考和研究. 本节通过回顾 FPH 框架与 PCANet，以及 PCANet 与经典的 CNN 之间的联系来初探这一问题.

早在 2005 年，Zhang 等人[69]已经注意到依次使用 Gabor 小波分解、局部二值编码与局部特征统计所生成的统计特征具有强大的表示能力与泛化能力. 这一特征处理过程被后来的研究者们[32,59,61,68]归纳为特征图生成、模式图编码和柱状图计算等三个步骤，本文称之为“FPH 框架”. FPH 框架遵从了

从低阶特征到高阶特征的逐层处理的过程：(1) 特征图用以分解和过滤原始像素特征；(2) 模式图用以重新组装各分解后的特征，形成局部纹理特征，同时也起到了对特征图压缩和对特征值规范化的作用；(3) 柱状图用以描述图像的空间结构信息.

继 Zhang 等人[69]的工作之后，FPH 框架被广泛应用于设计新的鲁棒人脸特征. 研究者们先后设计出了不同的特征图和模式图. 特征图主要包括：Gabor 幅值[69]、Gabor 相位[68]、单演信号[59]、梯度模[62]、梯度方向[63]等；模式图主要包括：局部二值模式(LBP)[84-85]、局部导数模式(LDP)[61]、象限比特编码(Quadrant Bit Coding，QBC)[68]、局部异或模式(Local XOR Pattern，LXP)[68]等. 不同的特征图需要与不同的模式图编码规则相匹配. 一般情况下，方向类(如梯度方向)或相位类(如 Gabor 相位)的特征图需要用象限比特编码[59,68,86]，而比特类的特征图需要使用局部异或模式编码[68]，幅值(如 Gabor 幅值或梯度模)或亮度类的特征图需要用局部二值模式或局部导数模式编码[62,69,84-85,87]. 自 2002 年 Ojala 等人[85]提出局部二值模式以来，模式编码的设计规则并没有发生实质性的变化，主体思路仍然是：首先通过与邻域特征值比较(如 LBP，LXP)或量化特征值(如 QBC)的方式产生比特流，然后将比特流编码为十进制. 因此，研究者们主要着眼于提取更好的特征图，以从输入图像中产生更多有用的特征，同时去除噪声的干扰.

PCANet[17]正是利用 CNN 对 FPH 框架中的特征图的生成过程进行改进而得到的，如图 9 所示. 与传统的特征图的生成过程不同的是：(1) PCANet 采用的滤波器组是从训练数据中学习到的，而非预先手工设计的；(2) PCANet 对输入图像进行了两层分解，而非单层分解.

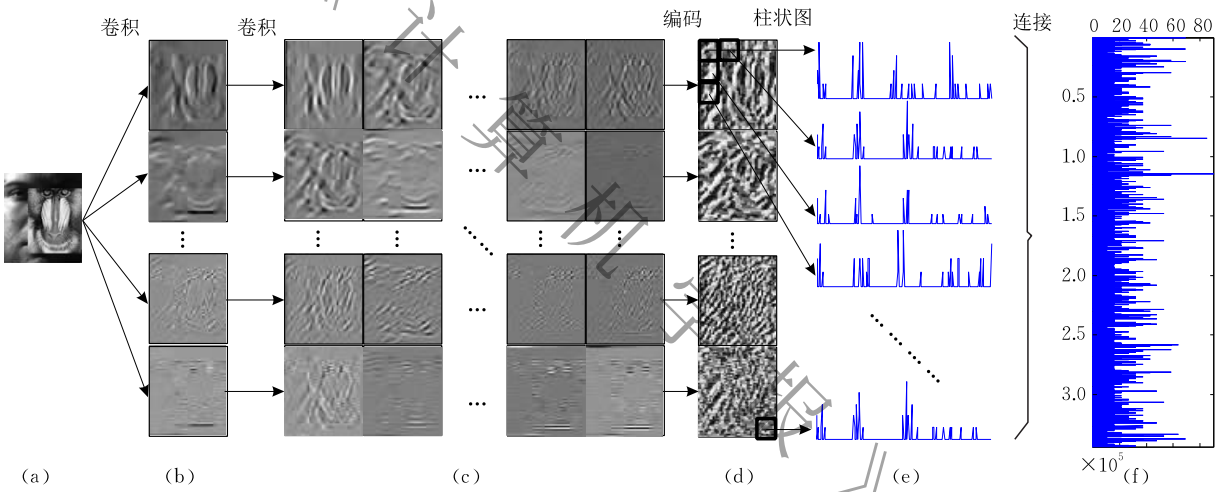


图 9 PCANet 的特征提取过程((a)～(c)特征图生成：输入图像的特征在各主方向上的两层分解；(d)模式图编码：对每组特征图进行二值编码，生成 8 个模式图；(e)～(f)计算并连接 8 个模式图的局部统计结构)

首先来看 PCANet 中滤波器组的训练过程. 这一过程可以看作是施加了空间局部约束的 PCA 的深度训练：(1) 将所有训练图像分割成大小相等的块；(2) 将所有图像块拉伸为列向量，组成一个数据矩阵；(3) 基于 PCA 从该数据矩阵中获取前 k 个主方向作为第一层滤波器组. 第 $l>1$ 层滤波器组的训练过程与此类似，只需将上述步骤(1)中的训练图像替换为用第 $l-1$ 层的滤波器组对 $l-1$ 层的训练图像的滤波结果即可. 与经典的深度网络最大的不同是，PCANet 的训练过程不需要通过随机梯度下降方法进行反向传播，这样做可以在一定程度上避免因训练数据较少而引发的过拟合问题.

图 10 对比了 PCANet 的两层滤波器组和经典的 Gabor 滤波器组. 如果说 Gabor 滤波器组很好地

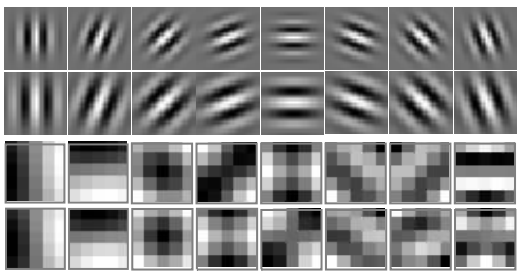


图 10 Gabor(上图)与 PCANet(下图)滤波器组的比较

模拟了人类视觉皮层 V1 视觉区域的感受野响应[88]，那么，PCANet 滤波器组显然并不具备这样的性质，然而，Chan 等人[17]的实验表明 PCANet 比 Gabor 特征具有更好的描述能力和鲁棒性. 究其原因主要在于 Gabor 滤波器组只是模拟了 V1 视觉区域的感受野响应，而 PCANet 则模拟了人类视觉皮

层对视觉信号的深层次处理过程，这再次表明了深度分解以及由此产生的高阶特征比浅层特征具有更好的表达能力和去噪能力. 图 11 从“奇点分解”的角度进一步解释了深度分解的意义：通过深度分解，既可以从输入图像的每一个像素点中“释放”出丰富的特征，又可以去除噪声的干扰.

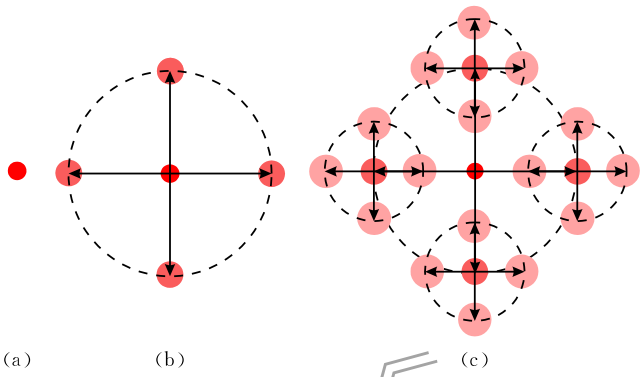


图 11　PCANet 的特征图生成过程的“奇点”解释((a)每个像素点及其邻域构成了一个“奇点”：包含着丰富的特征却不为人知；(b)～(c)奇点通过在每一个方向上的深度分解来释放自己的“能量”，每次分解，它的轮廓(特征)更加清晰，而所蕴含的能量却在衰减)

第 7 节的识别实验表明 PCANet 具有很强的遮挡鲁棒性. Chan 等人[17]猜想其原因在于：PCANet 的滤波器组能够检测并消除遮挡的影响. 然而，观察图 9(b)～(d)可以发现 PCANet 所产生的特征图和模式图仍然清晰地保留了遮挡的特征. 图 12 比较了 PCANet 响应对于同一张人脸图像在遮挡前后发生

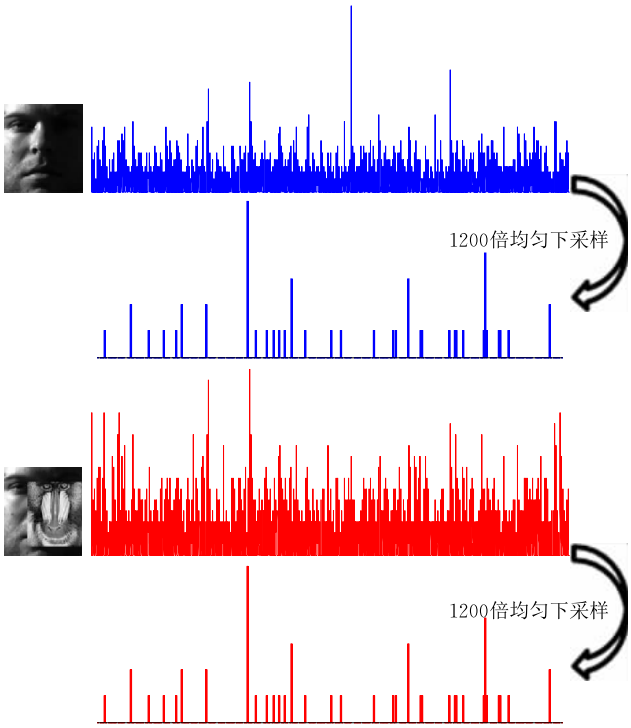


图 12　同一张人脸图像在遮挡发生前后的 PCANet 响应

的变化，可以发现：遮挡的存在使得 PCANet 响应变得更加“稠密”了，也就是说，PCANet 仍然花了很大代价来表示坏特征，然而，当进行了较大倍数(如 1200 倍)的均匀下采样后，PCANet 的响应变得惊人的一致. 这表明“稀疏性”仍然是特征学习应该遵循的一个基本准则. 如何基于稀疏性准则来提升 PCANet 响应的质量有待于进一步研究.

5.4　属性学习与卷积神经网络

人眼之所以可以迅速识别出一张有局部遮挡的人脸图像，一个主要原因是人眼能够“分解”并“理解”人脸图像中所包含的高阶属性，例如：头发的颜色、鼻子的轮廓等. 2009 年，研究者们[70,89-90]开始基于图像理解探讨基于高阶语义特征(即人类语言可描述的图像属性)的图像分解. Kumar 等人[70]建议将人脸图像分为人种、性别、年龄、脸型、头发颜色等 65 种属性，并提出基于低阶人脸特征(如颜色、边缘、梯度方向等)和支撑向量机来训练高阶属性的分类器. 然而，从低阶特征到高阶特征的变换可能远非常规的分类器所能模拟. 深度学习[78]通过从输入层到输出层的多层非线性映射以及基于反向传播的学习机制为解决这一问题提供了指导性方法. 2013 年，Luo 等人[71]基于“和积网络”描述各属性之间的高阶相关性，通过这种相关性可以推断那些因为遮挡而缺失的属性特征；2014 年，Zhang 等人[73]将局部表示与卷积神经网络相结合用以解决属性推断中的姿势变化问题. 上述方法都需要显式地标注属性，而属性标注需要耗费巨大的人力成本和时间成本[70].

那么，是否可以通过某种机制使学习到的特征自动区分图像中的各种属性呢？这一问题仍然可以从深度学习中寻求答案. 深度学习的理论研究表明深度网络具有强大的“分布式表达”能力[78]：深度网络的每个隐含层的每个神经元都可以视作一个属性分类器，每个神经元将图像空间一分为二，n 个神经元的组合就可以表达 2^n 个局部区域. 2015 年，Sun 等人[18]利用他们精心设计的 CNN 网络对人脸图像的神经元响应进行了系统分析，并指出：位于最高隐含层的神经元响应具有中等的稀疏性、对遮挡的鲁棒性和对高阶属性的选择性. 事实上，这种对高阶属性的选择能力(即神经网络的分布式表达能力)必然会带来神经元响应的判别性和稀疏性(例如，一个人不可能既是黑人又是白人)以及对遮挡的鲁棒性(图像的属性已经很好地分离，可以根据未遮挡的属性对图像进行分类).

为什么 Sun 等人[18]设计的 CNN 网络会具有这

样的分布式表达能力？这种能力是否可以拓展到现有的其它 CNN 网络？表 1 列出了近三年来所提出的用于人脸识别的主流 CNN 网络及其主要特征. 可以发现 CNN 网络的分布式表达能力主要可归因于如下三个因素：

首先，从低阶到高阶的层次性网络结构. 为了分离图像中的各种高阶属性，首先必须获取这些属性. 以 LeNet5[97] 为代表的 CNN 网络基于图像高阶属性的生成方式来设计网络结构：(1) 在较低的几层利用卷积从低阶特征(像素、纹理、方向等)逐层获取局部高阶特征. 由于图像中不同区域的局部高阶特征的形成过程基本类似，所以用卷积核共享权重；由于局部高阶特征的形成过程可能不是单一的，所以往往利用多重卷积；(2) 在较高的几层利用全连接进一步获取各局部高阶特征的相关性(因为大多数属性都不是独立存在的，如：年龄、人种等).

其次，有针对的分块训练. 主要表现在两个方面：(1) 加入局部连接层. 从表 1 可以发现，DeepFace[91-92] 和 DeepID 系列[18,64-65] 相对于 LeNet5[97] 在网络结构上最大的不同在于：都在卷积层和最高的全连接层之间加入了局部连接层. 这样做有助于针对局部特征进行专项训练，所训练的卷积核将对这一局部区域内的正常特征有较强的响应，而对异常特征(如遮挡)则会产生过滤的作用[92]；(2) 对输入图像进行分块训练. Sun 等人[18,64-65] 认为仅仅在网络中加入局部连接层仍然是不够的，还需要直接将输入图像划分为不同的局部区域进行专项训练. 这样做一方面可以增广训练样本，另一方面可以使得所训练的网络从一开始就能够较好地过滤局部异常特征.

最后，强力监督. 人脸图像的各种属性共同决定了其类别；反之，也可以由人脸图像的类别获取其各种属性. 由此可见类别信息对于属性分解至关重要，然而，研究者们在 CNN 网络的设计和训练过程中对这一问题的认识却并非是一步到位的. 2013 年，Sun 等人[98] 利用认证信号来训练网络(即网络的输入是成对的信号，而网络的输出是判别这对信号是否为同一类别)，在网络的最高隐含层获得了类内变化较小而类间变化较大的特征. 2014 年，Sun 等人[65] 和 Taigman 等人[91] 都意识到仅仅利用认证信号还不能充分强调类别之间的差异，提出利用识别信号来训练网络(即网络的输入是有类标的信号，网络的输出是信号的类标). 利用识别信号来训练网络要求有足够多的类别且每个类别有足够多的样本. 2014 年底，Sun 等人[64] 建议同时使用认证信号和识别信号来训练网络，以增强网络的判别性. 为了进一步加强对网络的监督，2015 年，Schroff 等人[93] 提出了 FaceNet，采用最小化三元组损失函数来训练网络. 同年，Parkhi 等人[95] 也基于最小化三元组损失训练 VGG 网络[99]. 与此不同的是，Sun 等人[18] 采用了更大规模的训练数据和更强力的监督训练(即对网络的每一层特征都同时施加认证和识别监督). 在如此强力的监督下，Sun 等人[18] 发现所训练的网络在最高隐含层形成的特征开始具有了中等的稀疏性、对遮挡的鲁棒性和对高阶属性的选择性.

表 1 近三年所提出的用于人脸识别的主流卷积神经网络及其主要特征(最后一列表示在 LFW 数据库上的认证准确率)

卷积神经网络	年份	特征维数	网络风格	网络层数	网络参数	网络个数	训练图像/类别	%
DeepFace[91]	2014	4096	加入了局部连接层的 LeNet	9	120 M	1	4.4 M/4030	97.35
DeepID[65]	2014	160×60×2	加入了局部连接层的 LeNet	9	~13 M	60	202 K/10 K	97.53
DeepID2[64]	2014	160×25	加入了局部连接层的 LeNet	9	~13 M	25	160 K/8 K	99.15
DeepID2+[18]	2015	512×25	加入了局部连接层的 LeNet	9	N/A	25	290 K/12 K	99.47
DeepFace+[92]	2015	1024	加入了局部连接层的 LeNet	9	N/A	1	500 M/10 M	98.37
FaceNet[93]	2015	128	Zeiler&Fergus GoogLeNet[94]	22 24	140 M 7.5 M	1	200 M/8 M	99.63
VGGFace[95]	2015	4096	深层的 LeNet	22/24/27	144 M	1	2.6 M/2622	98.95
LightenedCNN[96]	2015	256	引入了 MFM 激活的 LeNet	11 12	3.96 K 3.24 K	1	493 K/10 K	97.77 98.13
SparseConvNets[37]	2016	512	VGG	15	~4.6 M	25	290 K/12 K	99.55

从表 1 也可以看出近三年来 CNN 网络的一个发展趋势是：网络层数在加深，网络参数在减少. 从网络层数上来看，相对于 2014 年的 DeepFace 和 DeepID 所使用的 9 层网络，2015 年的 FaceNet[93] 和 VGGFace[95] 所使用的网络都超过了 20 层，而 DeepID2+[18] 虽然并没有增加层数，却增加了每一个卷积层的宽度(每一卷积层都加入了一个全连接层). 正如 Simonyan 等人[99] 所指出的，增加网络深度的一个重要意义在于：可以用更少的参数达到更强的表达能力. 然而，过于深层的网络会引发梯度消失/爆炸、性能退化等问题，He 等人[100] 对此进行了回顾，并给出了性能退化的解决方案. 可以预见的是未来的人脸识别将基于更深层次的网络结构. 2016 年，CNN 发展的另一个趋势是：减少网络的参

数. 为了用有限的训练样本训练出性能更好的神经网络，Sun 等人[37]提出了 SparseConvNets；为了提高计算性能和减少存储空间，Wu 等人[96]提出了轻量级 CNN.

更深层次和更轻量级的网络对有遮挡人脸图像的识别性能有着怎样的影响有待于进一步研究.

6　鲁棒分类器中的鲁棒特征嵌入

现有的鲁棒分类方法，如第 2 节和 3.2 节所提到的鲁棒回归方法和鲁棒误差编码方法，大都是基于像素域设计的. 为了进一步提升鲁棒分类器的性能，一种直接的思路是将鲁棒特征直接嵌入到鲁棒分类器中，如 2.2 节提到的 GRRC[13,25]就是将 Gabor 特征直接嵌入 CRC 而得到的. 直接将鲁棒特征嵌入到鲁棒分类器尽管一般也能在一定程度上提升识别性能，但同时也会引发一些新的问题：基于像素域设计的鲁棒分类器由于没有考虑到新的特征所具有的新的性能而无法最大限度地提升识别性能，如本文实验部分的图 21 和图 22 所示. 那么，该如何将鲁棒特征有效地嵌入鲁棒分类器中呢？Yang 等人[32]和 Liang 等人[19]的工作给出了解决这一问题的范例.

2013 年，Yang 等人[32]将统计局部特征(Statistical Local Feature，SLF)嵌入到 RSC[14]中，提出了基于鲁棒核表示(Robust Kernel Representation，RKR)的分类方法. SLF 是基于 FPH 框架的统计性特征. 对于局部柱状图特征，现有的研究[84]表明直方图交叉和卡方距离是比 ℓ_1 或 ℓ_2 范数更有效的距离度量方法. 因此，将柱状图特征嵌入鲁棒分类器的问题归根结底是将直方图交叉核或卡方核嵌入鲁棒分类器的问题. RKR[32]是将直方图交叉核嵌入到 RSC 的结果. Gao 等人[101]和 Wang 等人[102]进一步讨论了如何将核方法嵌入一般的子空间回归方法. 由于 PCANet 也是一种局部柱状图特征，Chan 等人[17]建议将卡方距离嵌入到最近邻分类器对 PCANet 特征分类.

2015 年，Liang 等人[19]将 IGO 特征 $\mathcal{G}(\cdot)$ 嵌入到 SSEC[9]中，提出了混合误差编码模型(Mixed Error Coding，MEC). 在 SSEC 中最关键的步骤是通过对重构误差的阈值聚类和结构聚类进行遮挡检测. 而在 IGO 域中，这样的遮挡检测是隐含的(详见 5.2 节). 为了显式检测遮挡位置，Liang 等人[19]利用 IGO 域中图像之间的相似性度量式(16)来度量图像的各像素点之间的相似性：将式(16)收缩到像素点的 $b\times b$ 邻域范围，用该邻域范围内的小块图像之间的相似度作为像素点之间的相似度，进而，就可以得到两幅图像之间的所有像素点的相似度 $\widetilde{\mathcal{S}}$(称为结构化相似度)，$\widetilde{\mathcal{S}}$ 中那些相似度较小的像素点就可以作为异常特征点被检测出来. 为了进一步准确获取遮挡的位置，MEC 还须对结构误差 $\tilde{e}=1-\widetilde{\mathcal{S}}$ 进行阈值聚类和结构化聚类. 显然，结构误差度量 $\tilde{e}$ 并不适合于恢复重构系数 x，MEC 采用了与 SSEC 类似的方法，即基于最小 ℓ_2 回归计算 x. 由于 MEC 在计算重构系数 x 和结构误差 $\tilde{e}$ 时采用了不同的误差度量，因此，MEC 是一种混合误差编码模型.

7　仿真与实验

本节通过实验来验证前面提到的主流的鲁棒特征和鲁棒分类器对有遮挡人脸识别的有效性. 表 2 列举了本节实验所用到的六种特征，其中，IGO 特征和 Weber 特征分别指经过了 $\mathcal{G}(\cdot)$ 变换(式(17))和 $\mathcal{W}(\cdot)$ 变换(式(21))的 IGO 特征和 Weber 特征. 这六种特征大体上可以分为三类：浅层特征(Pixel/IGO/Weber/Gabor)、由浅层向深度过度的 FPH 特征(SLF/PCANet)、深度 CNN 特征(DeepID/VGG/LCNN). Tzimiropoulos 等人[33]和 Yang 等人[32]的实验表明 Gabor 特征对有遮挡人脸识别的性能弱于 IGO 特征和 SLF 特征. 因此，对于浅层特征，我们重点关注 Pixel、IGO 和 Weber，Gabor 特征主要用于构造遮挡字典以及与之相匹配的 GRRC 分类器[13,25].

表 2　本节实验所用到的鲁棒特征

缩写	全称
Pixel	原始像素特征
LBP[84,87]	局部二值模式
IGO[33]	图像梯度方向
Weber[81]	Weber 脸
Gabor[13,25]	基于 Gabor 变换的特征
SLF[32]	统计局部特征
PCANet[17]	基于 PCANet 的特征
DeepID[65]	深度隐含层类别特征
VGG[95,99]	牛津大学视觉几何组开发的深层 CNN
LCNN[96]	轻量级 CNN

表 3 列举了本节实验所用到的 10 个分类器，其中，NN 分类器专指基于卡方距离度量的最近邻分类器，仅用于对 PCANet 特征分类；NS/LRC 分类器由于没有采用协同表示机制，在很多情况下性能不够鲁棒(详见 3.1 节)，仅在 7.2 节中用于测试遮挡所造成的数据损失对朴素分类器的影响. 根据对某种特定特征的依赖程度，可将表 3 的分类器分为两类：非定制的分类器，包括 CRC、CESR、RSC、

RNR 和 SSEC，它们与具体的特征无关；定制的分类器，包括 MEC、GRRC、RKR 和 NN，它们是分别针对 IGO、Gabor、SLF 和 PCANet 设计的分类器. 根据是否专门针对遮挡问题或对遮挡先验信息的利用程度，可将表 3 中的分类器分为三类：通用分类器，即完全没有利用遮挡先验的分类器，包括 NN、CRC 和 RNR；弱遮挡假设分类器，即只做了一般性噪声假设的分类器，包括 CESR、RSC 和 RKR，此类分类器通常都采用了对输入信号加权的方式，因此也称为“带权分类器”；强遮挡假设分类器，即需要显式检测遮挡或对遮挡信息明确编码的分类器，包括 SSEC、MEC 和 GRRC. 为了兼顾效率与公平，所有需要稀疏重构的分类器（除了 NN 和 NS 外）均采用非负最小二乘计算回归计算重构系数 x，但基于 Gabor 特征的分类器（如 GRRC 或 CRC＋Gabor 等）除外，实验表明，Gabor 特征更适合用最小二乘计算重构系数；CRC 的正则化参数 λ 取 0.001，所有其它分类器和特征提取方法均采用作者在原文中提供的默认参数.

表 3 本节实验所用到的鲁棒分类器

缩写	全称
NN	最近邻分类器
NS/LRC[103]	最近子空间分类器/线性回归分类器
CRC[24]	基于协同表示的分类器
CESR[10]	基于相关熵稀疏表示的分类器
RSC[14]	鲁棒稀疏编码
RNR[29]	基于鲁棒核范数正则化回归的分类器
SSEC[9]	结构化稀疏误差编码
MEC[19]	混合误差编码
GRRC[13,25]	基于 Gabor 特征的鲁棒表示与分类方法
RKR[32]	基于鲁棒核表示的分类器

7.1 CNN 网络的训练与精调

我们选择 PCANet[17]、DeepID[65]①、VGG[95] 和 LCNN[96]② 等四种经典的深度网络特征用于对比实验. 可以认为 PCANet 是传统的特征提取方法向 CNN 网络的过渡，而 DeepID、VGG 和 LCNN 是三种有代表性的 CNN 网络，分别代表了联合的多元深度网络、加深的神经网络和轻量级神经网络.

DeepID 的训练. PCANet、VGG 和 LCNN 的作者公布了他们训练好的网络模型以及模型训练所采用的数据集，可以免费下载③；而 DeepID 系列的网络模型至今尚未公开，我们选取 Caffe 框架[104] 和 CelebA[105]④ 数据集（图 13 第 1 行）训练 DeepID 的网络模型. Sun 等人[65] 建议基于 25 个图像区域并发训练 DeepID 的 25 个网络模型，以同时达到增广训练数据集和增强网络模型的适应性的目的. 受实验条件限制，本文基于 5 个人脸区域（图 13 第 2 行）及其水平翻转，训练 10 个 DeepID 网络模型，所获得的滤波器核样例如图 13 的第 3 行所示. 可以看出，相对于基于整张人脸图像训练所得的滤波器核，基于四个局部区域训练所得的滤波器核具有较为明显的结构特征，并且不同局部区域的个别滤波器的结构具有一定的相似性，这表明人脸图像的各个局部区域具有共享的统计结构[78]. 为了确保所训练的 DeepID[65] 模型的有效性，我们在 LFW 数据库[106]（详见 7.4 节）上进行人脸认证实验，取得了接近于作者原文的准确率（96.13%）.

图 13 来自于 CelebA 的人脸图像（第 1 行）、本文用于训练 DeepID 的人脸图像区域（第 2 行）、训练所得的滤波器核样例（第 3 行）

本文在受控（见 7.2 和 7.3 节）和非受控（见 7.4 节）两种环境下验证深度网络模型对于有遮挡人脸识别的有效性. 值得注意的是深度网络模型通常都是在大规模的非受控环境下训练得到的，如果直接将其应用于受控环境，往往需要进一步对网络模型参数精调. 然而，精调网络模型仍然需要较大规模的训练集. 当目标数据集上的训练数据非常少，甚至目标数据集的测试数据与训练数据发生严重偏移时，在目标训练集上的精调必然会导致过拟合，且过拟合的程度会随着训练迭代次数的增加而增大. Ghazi 等人[107] 也报告了类似的实验并现象. 该如何将大规模非受控环境下训练得到的深度网络模型应用于小规模非受控环境下的目标数据集是尚待解决的问题. 本节接下来的实验将直接使用非受控环境下训练得到的 CNN 网络模型（DeepID/VGG/LCNN），不再针对每个数据库进行精调. 由于所采用的 PCANet 是在非受控环境训练的，因此，在 7.2 节～7.4 节的实验中我们将看到 PCANet 与其它三个卷积网络模型恰好相反：它在受控环境下的识别性能良好，而在非受控环境的识别性能变差.

① DeepID 系列实际上有四个版本[3,18,64-65]，每一个版本的性能提升实际上都是比较微小的，但网络模型的复杂性却有很大提升，本文选择 DeepID 最初的版本[65].
② Wu 等人[96] 提供了 A 和 B 两个版本的 LCNN，本文选择性能更好的 LCNN-B 网络.
③ https://github.com/betars/Face-Resources
④ http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

7.2 基于 Extended Yale B 数据库的有光照变化与模拟遮挡的人脸识别

本节基于 Extended Yale B[108-109] 人脸数据库测试主流的分类器(见表 3)在不同的特征域(见表 2)中对不同程度的光照变化和不同程度的遮挡的承受能力. Extended Yale B 数据库包含了 38 个人的具有不同光照变化的脸部图像，按照光照变化条件可划分为五个子集：I(共 262 张)、II(共 455 张)、III(共 453 张)、IV(共 524 张)、V(共 712 张)，从子集 I 到子集 V，光照逐渐变暗，如图 14 的第 1 列所示. 为了模拟不同程度的光照和不同程度(百分比)的遮挡的混合情况，我们在子集 III、IV 和 V 的每张图像中分别添加 0～90%(以 10%为间隔)的模拟遮挡(狒狒)，且遮挡的位置是随机的，如图 14(c)～图 14(e)所示，最终得到的子集 III、IV 和 V 中共有(453＋524＋712)×10＝16 890张包含了不同程度的光照变化和遮挡的图像. 选取子集 I 和子集 II 的所有图像作为训练集，选取扩充后的子集 III、IV 和 V 分别作为测试集 I、II 和 III. 所有的训练集和测试集图像均被裁剪为 96×84 的像素矩阵.

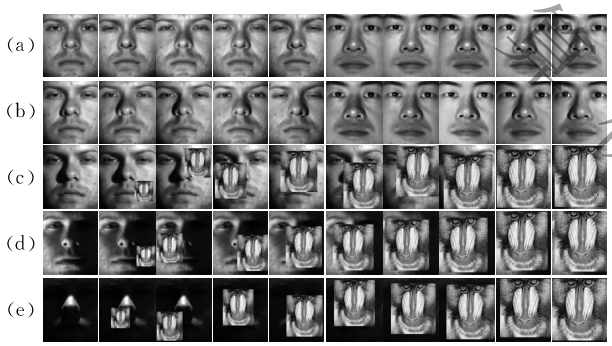


图 14　本节实验所采用的来自于 Extended Yale B[108-109] 的训练数据和测试数据样例((a)～(b)为训练样本：由子集 I 和 II 组成，不含遮挡且只有轻微的光照变化；(c)～(e)为测试样本：分别由子集 III、IV、V 组成，并加入了 0～90%的模拟遮挡(狒狒)，且遮挡位置是任意的)

首先通过一个简单的实验(图 15)来分析遮挡导致识别性能下降的主要原因. 图 15 表明：如果能够彻底消除遮挡的影响，即使只使用“浅层”的鲁棒特征(IGO 和 Weber)和最简单的 NS 分类器，也能达到良好的识别效果；相反，如果不能完全排除遮挡的影响，即使使用“深层”的 PCANet 特征以及与之相匹配的 NN 分类器，当遮挡比例增加到一定程度时，识别率也会急剧下降. 这一方面说明了检测并排除遮挡的重要性，另一方面也说明了遮挡本身的存在比遮挡所造成的特征损失更容易导致识别性能的下降，而深度特征(如 PCANet)的“遮挡不变”与“光照不变”仍然有其局限性.

四种浅层特征的鲁棒性分析. 从图 15 中也可以发现：在只有光照变化的情形下(遮挡已经完全排除)，四种浅层特征的鲁棒性排序为：Weber＞IGO＞LBP＞Pixel，并且它们的性能差异随着光照条件的变坏而更加明显. 那么，在有遮挡的情形下，随着遮挡比例的增加，这四种浅层特征是否仍有类似的规律？图 16 对此进行了验证，可以发现，对于绝大多数分类器而言，这一规律仍然成立. 另外，从图 16 也可以看出，鲁棒特征与鲁棒分类器相辅相成，彼此突出：随着识别难度的增加(光照条件恶化和遮挡比例增大)，在像素域中，分类器的鲁棒性差异越来越不明显，而在鲁棒特征(IGO/Weber)域中，分类器的鲁棒性差异却越来明显.

三种 CNN 特征的鲁棒性分析. 图 17 基于通用分类器 CRC 和带权分类器 CESR 比较了三种 CNN 特征的识别性能，可以发现，其鲁棒性排序为：DeepID＞LCNN＞VGG，并且在 CNN 特征域中，带权分类器(如 CESR)相对于通用分类器(如 CRC)的性能优势

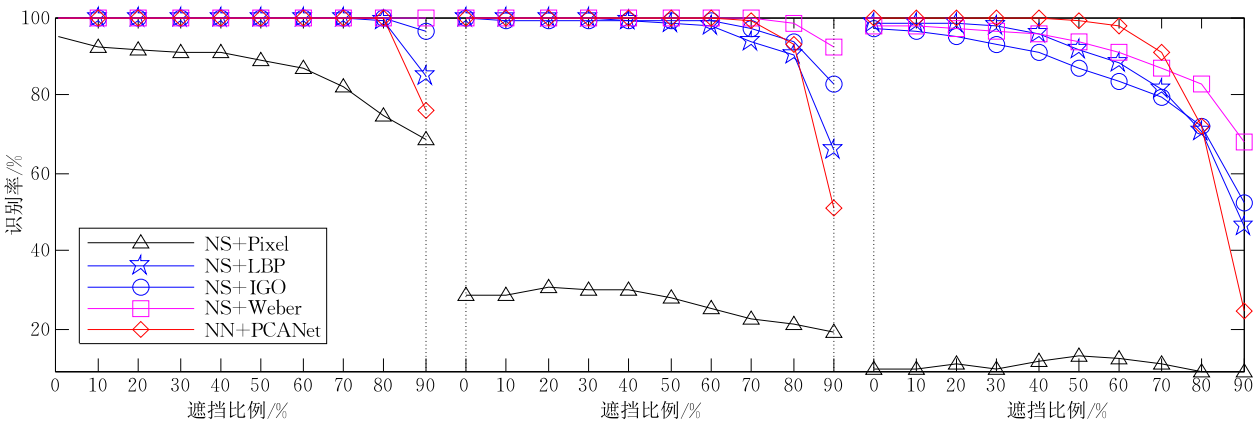


图 15　遮挡所造成的局部特征损失对不同的光照子集(从左至右依次为光照子集 III、IV、V)的识别率的影响(NS 分类器已知遮挡支撑并在识别过程中丢掉了被遮挡区域的特征；NN 分类器不知道遮挡支撑，在识别过程中用到了整幅图像的 PCANet 特征)

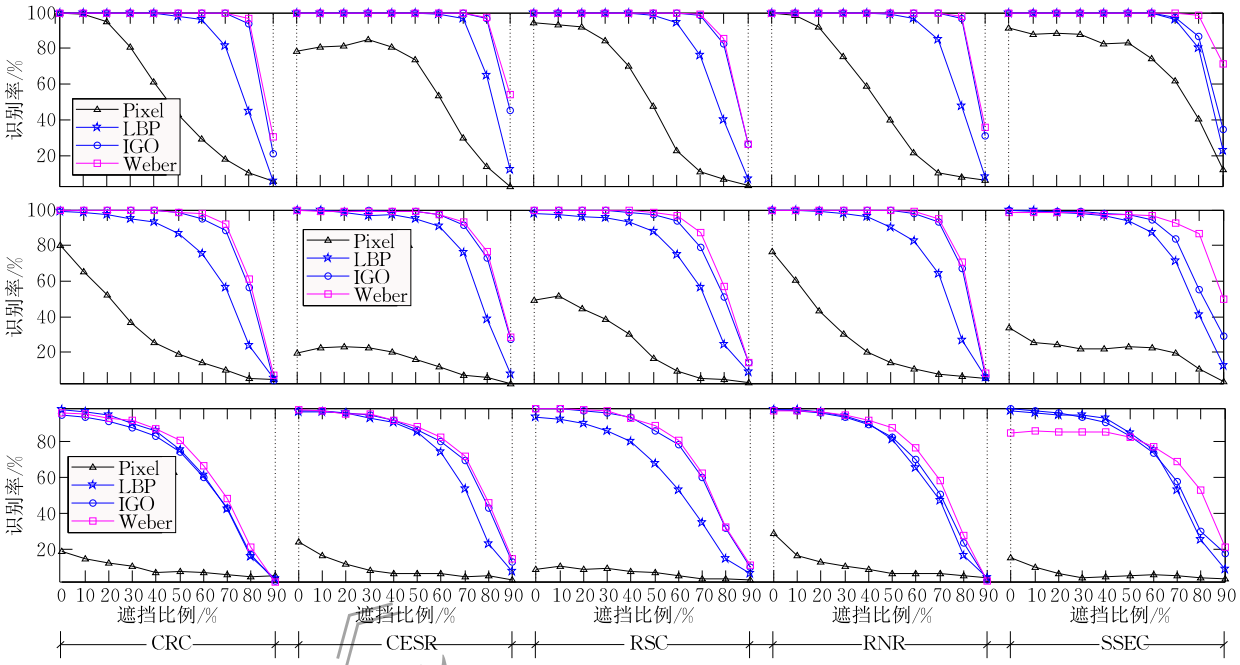


图 16 五个非定制分类器(CRC/CESR/RSC/RNR/SSEC)使用四种浅层特征(Pixel/LBP/IGO/Weber)在 Extended Yale B 的三个测试集(第 1 至 3 行依次对应于测试集 I、II 和 III)上的识别率

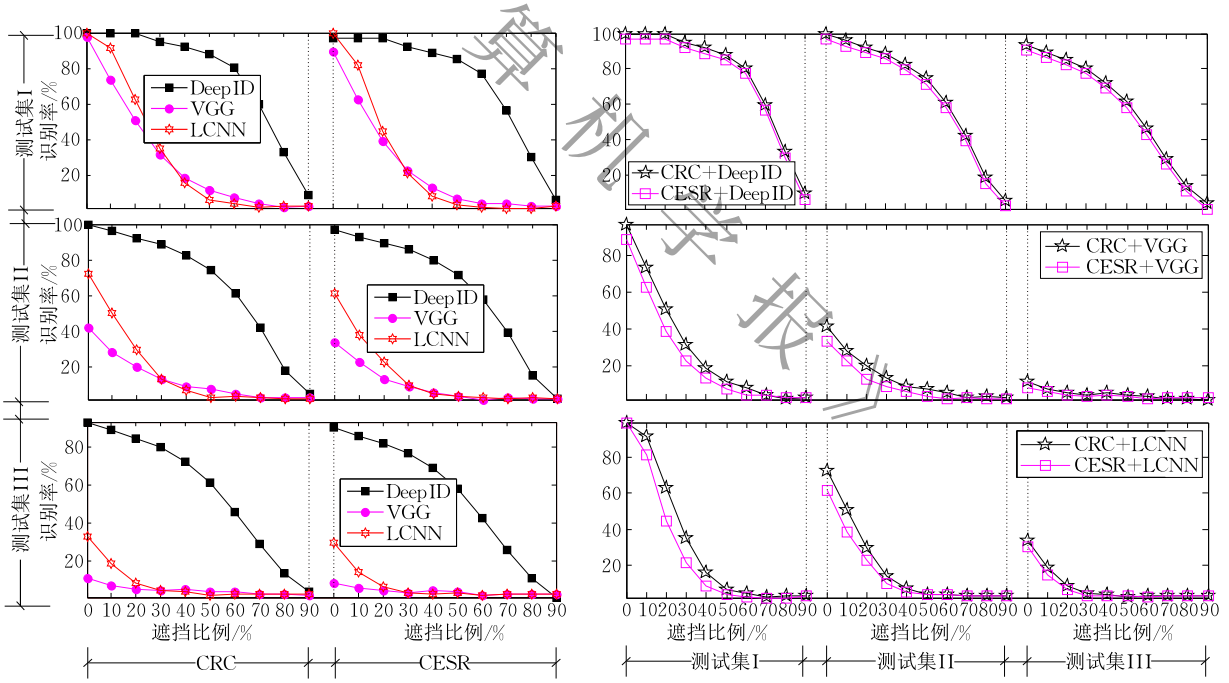


图 17 通用分类器 CRC 和带权分类器 CESR 使用三种 CNN 特征在 Extended Yale B 的三个测试集上的识别率

并不像其在浅层特征域中那样明显.鲁棒分类器的性能分析.综合对比图 16 和图 17,可以发现:通用分类器 RNR 和 CRC 更擅长处理光照问题(即遮挡比例较小的情形),这种能力在像素域中可以更为清晰地看出,但对遮挡的鲁棒性相对较弱;SSEC 能够很好地利用 Weber 特征,但在光照条件恶化的情况下,容易对遮挡过拟合(把部分极端的光照变化也当成了遮挡),从而导致其识别性能在遮挡比例较低的情形下反而不如通用分类器 CRC;CESR 在 IGO 和 Weber 域中都具有良好的识别性能,而同样基于迭代重权模型的 RSC 的性能相对较弱(详见 4.1 节的分析);尽管 MEC 是专门针对 IGO 定制的鲁棒分类器,但 CESR 在 IGO 域中的识别性能与 MEC 相当,这主要是由于"狒狒"遮挡仅存在于测试集中,仅通过噪声抑制(而不是消除)就能获取很好的识别性能.

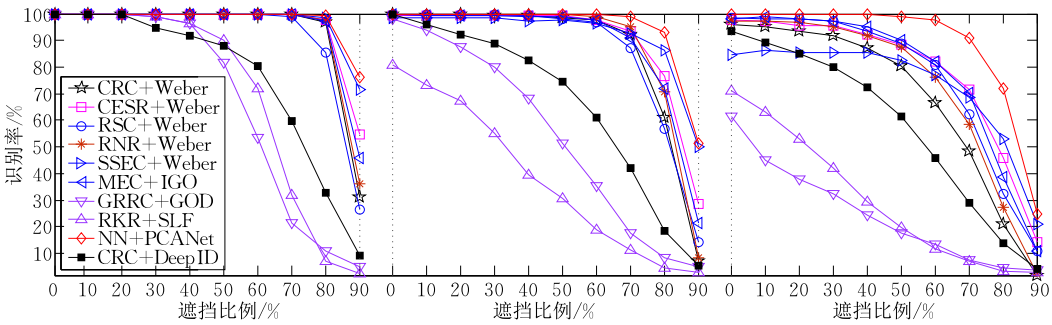


图 18　使用不同分类器和不同特征的最优组合在 Extended Yale B 的三个测试集上的识别率(从左至右依次为测试集 I、II、III)

由上述分析可知，鲁棒分类器与鲁棒特征存在某种程度上的最优组合. 图 18 对此进行了分析. 基于 FPH 框架的 NN＋PCANet 在各个测试集上取得了最优的识别性能，表 4 给出了具体的识别率数值；同样基于 FPH 框架的 RKR＋SLF 在三个测试集上的识别率反而几乎都是最低的，这再次表明图像的深度分解对于提升特征鲁棒性的重要作用；GRRC＋GOD 的识别性能基本上是最弱的，这表明基于固定形式训练的遮挡字典对现实遮挡的泛化能力较弱，难以从待识别图像中有效分离出遮挡成分；基于三种 CNN 特征的 CRC 分类器的性能远远弱于 NN＋PCANet，具体原因详见第 7 节的分析；SSEC＋Weber 在遮挡比例较大的情形下识别率最接近于 NN＋PCANet，但在极端光照条件下容易对遮挡过拟合；MEC＋IGO 尽管充分利用了 IGO 的特性，但其识别性能在总体上弱于 CESR＋Weber，这表明 Weber 特征对光照和遮挡具有更强的鲁棒性；与图 15 中的完全排除了遮挡影响的 NS 分类器的识别性能相比，SSEC＋Weber 与 MEC＋IGO 尽管充分地融合了鲁棒特征与遮挡检测，但仍然未能完全排除遮挡的影响，这表明遮挡检测仍然具有挑战性.

表 4　Extended Yale B 上的最优识别率/%，也是 NN＋PCANet 的识别率(I、II 和 III 分别表示测试集 I、II 和 III)

	0～30%	40%	50%	60%	70%	80%	90%
I	100	100	100	100	100	99.56	76.26
II	100	100	100	100	99.05	93.16	51.14
III	100	99.72	98.88	97.62	90.76	71.85	24.79

7.3　基于 AR 数据库的有强光光照和两类实际遮挡的人脸识别

本节在不同的像素维下针对有实际遮挡的人脸图像测试各鲁棒分类器和特征的识别性能. 一般来说，随着像素维的增高，识别算法的性能也会随之增强，但当像素维高到一定程度时，数据中的噪声可能会造成较大的干扰，从而导致识别性能的下降，我们将这一现象称为“识别性能退化”. 性能退化的程度可以作为评价特征和分类器的鲁棒性的一种指标.

本节基于 AR[36] 人脸数据库进行识别实验. AR 库中包含了 126 个人的 3276 张图像，平均每人 26 张，分别在两个不同的时间段内采集，包含了不同的表情、光照和遮挡等变化. 我们按照如下方式部署实验环境. 从 AR 库中选取 119 个人(65 个男性和 54 个女性)的人脸图像作为实验数据. 训练集选取 119×8＝952 张不含遮挡和光照变化的人脸图像，如图 19(a)所示；测试集 I 选取 119×3＝357 张有左侧强光、右侧强光和正面强光等三种光照变化的人脸图像，如图 19(b)所示；测试集 II 选取 357 张有太阳镜遮挡且混合了正常光照、左侧强光和右侧强光等三种光照条件的人脸图像，如图 19(c)所示；测试集 III 选取 357 张有围巾遮挡且分别包含了正常光照、左侧强光和右侧强光等三种光照条件的人脸图像，如图 19(d)所示. 所有图像首先被裁剪和对齐为 112×92 的像素矩阵. 为了系统测试现有方法在不同维度下的识别性能，对所有图像进行了 1 倍、2 倍、4 倍和 8 倍下采样，对应的维数分别为 112×92＝10 304、56×46＝2576、28×23＝644 和 14×11＝154.

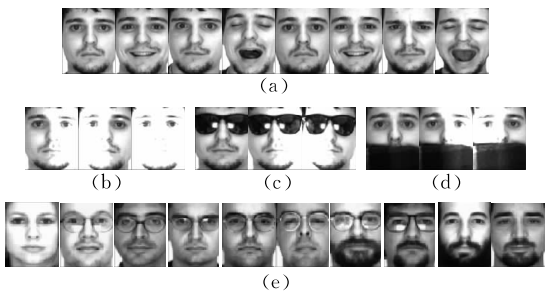


图 19　本节实验所采用的来自于 AR[36] 库的训练数据和测试数据样例((a)训练集：不含遮挡且具有正常光照；(b)测试集 I：仅包含光照变化；(c)测试集 II：有太阳镜遮挡和光照变化；(d)测试集 III：有围巾遮挡和光照变化；(e)容易与测试样本相混叠的训练样本)

本节实验所用数据比 7.2 节所用数据更有挑战性：(1) 实际遮挡比模拟遮挡更容易引发“局部混叠”，7.2 节中所用到的遮挡是人为添加的“狒狒”图像块，在训练集中基本不存在与“狒狒”具有较强相干性的特征，而本节实验的测试集中尽管只用到了两种实物遮挡，但因此而引发的“混叠现象”大量存在，如测试集中的黑色围巾和太阳镜等容易与训练集中的黑色胡子和普通框架眼镜(图 19(e))交互混叠，尤其是戴有普通框架眼镜的人大量存在于训练集中，很容易与测试集 II 中的太阳镜遮挡混叠，这就导致对测试集 II 的识别成为本节最具挑战性的任务；(2) 本节测试集中的光照变化主要是由强光光照引起的，与 7.2 节的测试集中所普遍存在的弱光光照不同，强光光照所引起的数据损失难以用常规的方法补偿，实际上也形成了一种遮挡，我们称之为“强光遮挡”. 与实物遮挡不同的是，由光照所引发的遮挡一般在局部范围内的像素值变化非常缓慢(属于低频遮挡)但通常面积比较大，如图 19(b)中的正面强光所引发的数据损失在 80% 以上. 当强光遮挡和实体遮挡相混合时(如图 19(c)和图 19(d))，图像识别被普遍认为是更有挑战性的任务[19,33].

首先比较四种常用的浅层特征(Pixel/LBP/IGO/Weber)的鲁棒性及四个常用的分类器(两个通用分类器 CRC 和 RNR，两个带权分类器 CESR 和 RSC)的识别性能，如图 20 所示. 可以看出，在测试集 II 上的识别任务更有挑战性，这与我们之前的分析一致. 对于各分类器，这四种浅层特征的性能总体上仍然满足：Weber＞IGO＞LBP＞Pixel，类似于图 16 的统计结果. 关于各分类器可以得到如下结论：(1) 就带权分类器 CESR 和 RSC 而言，在 Pixel 域中，RSC 最优；而在鲁棒特征域 IGO/Weber 中，CESR 几乎总是优于 RSC. 这一结论与 7.2 节中的结论类似，是由 CESR 和 RSC 所采用的优化算法不同造成的(详见 4.2 节)；(2) 通用分类器与带权分类器相比，随着所使用特征的鲁棒性的增强，通用分类器在低维情形下与带权分类器的性能差距逐渐缩小，而在高维情形下逐渐超越了带权分类器，尤其是 RNR 在 Weber 域中取得了较为显著的优势. 图 21(a)和(b)以 CESR 和 RNR 在 Weber 域中的识别过程为例说明了产生这一现象的原因：带权分类器容易为特征图中特征值较小的特征点赋以较高的权值；(3) 各分类器的性能退化问题在 IGO 域中最为严重，在 LBP 和 Weber 域中次之，在 Pixel 域中基本不存在；带权分类器 CESR/RSC 的退化现象比通用分类器 CRC/RNR 更严重. 图 22(a)以 CESR 在 IGO 域

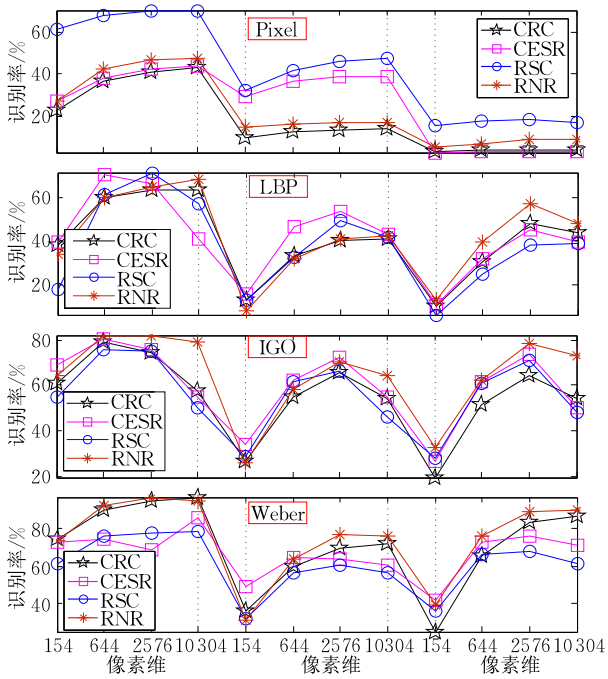


图 20 通用分类器(CRC/RNR)和带权分类器(CESR/RSC)使用四种浅层特征(Pixel/LBP/IGO/Weber)在 AR 的三个测试集上的识别率(从左至右依次为测试集 I、II、III)

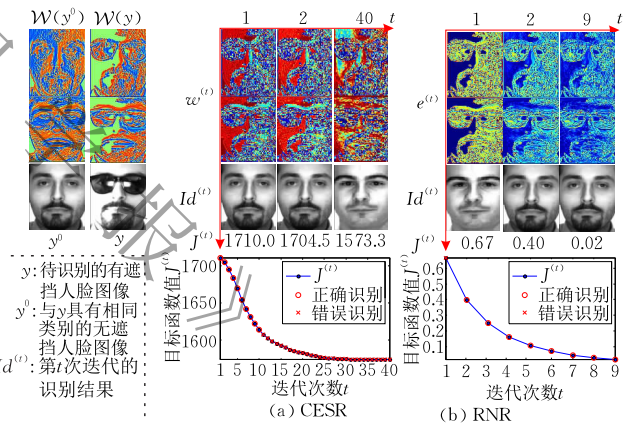


图 21 基于误差编码的乘法模型(CESR)和加法模型(RNR)在鲁棒特征域(Weber)中识别有遮挡图像 y((a)、(b)的上图显示在初始迭代时，CESR 和 RNR 都在右侧脸的高亮区域产生了较小的误差，但 CESR 基于乘法模型抑制噪声，为该区域赋予了较高的权值，导致其在随后的迭代过程中趋向于匹配右侧具有高亮特征的人脸图像；而基于加法模型的 RNR，不会因此对后续迭代产生太大影响；(a)、(b)的下图显示 CESR/RNR 的目标函数随着迭代次数的变化以及识别的正确与否：收敛并不一定意味着正确识别)

中的识别过程为例说明了这一现象产生的原因.

在第 6 节中，我们说明了鲁棒特征往往还需要辅之以定制的鲁棒分类器才能最大限度地发挥其作用，图 20 显示直接将鲁棒特征 IGO 和 Weber 嵌入到鲁棒分类器 CESR 和 RSC 不仅容易导致性能退化，而且其识别性能往往还不如通用分类器 CRC 和

RNR. 图 23 针对第 6 节中提到的 IGO、Gabor、SLF 和 PCANet 等特征分别测试了为之定制的 MEC、GRRC、RKR 和 NN 等鲁棒分类器在三个测试集上的识别性能. 可以看出定制的分类器的确在不同程度上提升了直接将鲁棒特征嵌入相应的基础分类器的性能，值得注意的是如下几点：(1) MEC 基本上避免了 IGO 特征的性能退化问题，图 22(b)和(c)通过比较 MEC 及其基础分类器 SSEC 分析了 MEC 的工作原理；但在最具有挑战性的测试集 II 上，MEC 仍然出现了性能退化；(2) 尽管 Gabor 特征能够很好地模拟人类视觉皮层的响应，但不论是 GRRC+GOD 还是 CRC+Gabor 的识别性能都要弱于 CRC+Weber(与图 18 中的统计结果一致). 图 24(上图)对此进行了分析；(3) 在只有强光遮挡的测试集 I 上，RKR+SLF 与 NN+PCANet 都不占优势；而在有遮挡的测试集 II 和 III 上，NN+PCANet 与 RKR+SLF 的性能优势逐渐凸显，尤其是 NN+PCANet 的性能优势尤为明显(最高维除外). 这表明了 FPH 特征更善于处理有纹理和方向等高频变化的图像，但不善于处理有大面积的低频变化的图像. 图 25 对这一现象进行了分析.

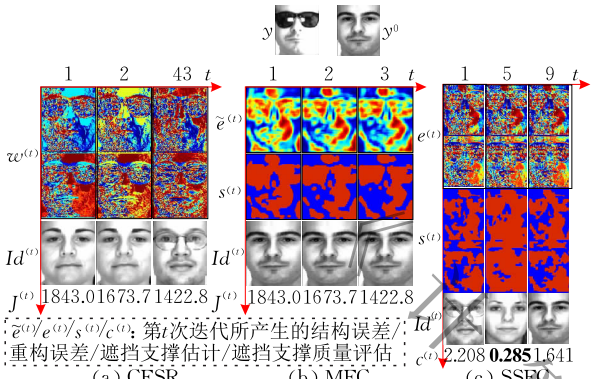


图 22　CESR、MEC 和 SSEC 在 IGO 域中识别有遮挡图像 y 的例子((a) CESR 的性能退化：特征值为 0 的特征点容易被 CESR 赋以较高的权值，这种偏执在较高的维度上被进一步放大了；(b)、(c) MEC/SSEC 的识别过程. 由于利用了 IGO 的“零和差异”进行结构误差度量，MEC 的结构误差 $\tilde{e}^{(t)}$ 比 SSEC 的重构误差 $e^{(t)}$ 具有更好的空间局部性和边缘光滑性；(c)中显示，虽然 SSEC 在最后一次迭代正确识别了 y，但是，由于在第 5 次迭代时发生了对遮挡的过拟合，导致其遮挡支撑评估策略(式(15))失效，最终识别错误)

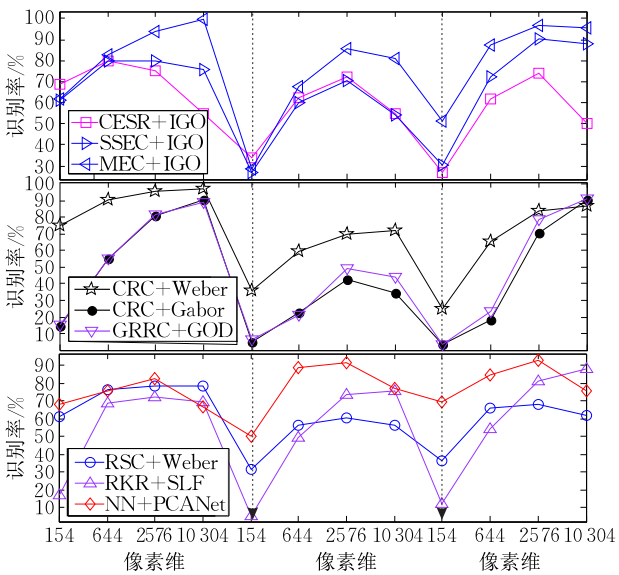


图 23　各定制分类器(MEC/GRRC/RKR/NN)及与之紧密相关的(分类器+特征)在 AR 上的识别率(从左至右依次为测试集 I、II、III)

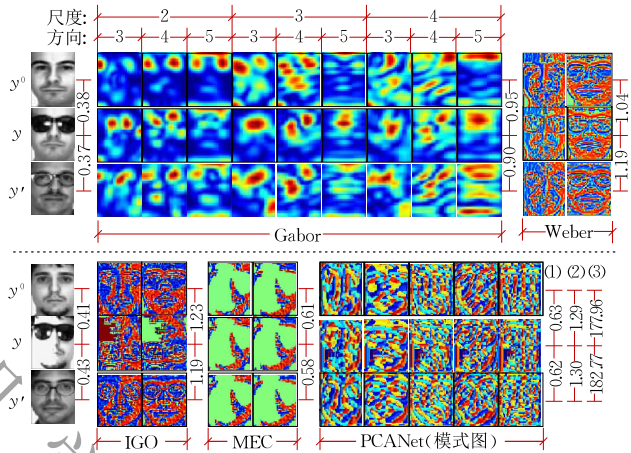


图 24　像素域及变换域中的混叠现象(y^0 与待识别图像 y 具有相同类别，y' 为容易与 y 发生混叠的训练图像(y^0 和 y)以及(y' 和 y)之间的 ℓ_2 距离标注在其右侧，所有特征图都进行了 ℓ_2 规范化. 从类内和类间的 ℓ_2 距离来看：上图中的 Gabor 变换放大了像素域中的局部混叠，而 Weber 特征则消除了混叠；下图中的 IGO 特征放大了像素域中的局部混叠，MEC 虽然可以去除遮挡和光照的影响，但仍然不能消除混叠；PCANet 的模式图右侧的数字(1)～(3)所标注的距离分别表示其模式图的 ℓ_2 距离、柱状图的 ℓ_2 距离、柱状图的卡方距离. PCANet 的模式图未能消除混叠，而其柱状图消除了混叠)

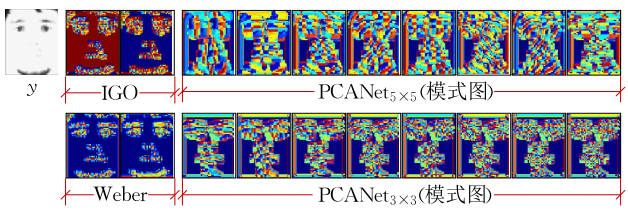


图 25　具有强光遮挡的图像 y 在 IGO、Weber 和 PCANet 域中的特征图及其鲁棒性分析(PCANet$_{b\times b}$ 表示采用 $b\times b$ 的卷积核进行深度滤波和模式编码后得到的模式图. 在 IGO 域中，y 的强光区域呈现出了两种相反的特征，这会导致以内积运算为主的线性分类器如 CRC 和 RNR 不能全部消除强光的影响；而 Weber 特征可以在很大程度上避免这一问题，其水平方向和垂直方向上的特征子图都将原始图像的强光区域置为了 0. PCANet 的模式图中呈现出了高频区域(眼睛、鼻子等)的特征向低频区域(强光区域)扩散的现象)

图 26 比较了三种 CNN 特征对识别性能的影响，可以得到与图 17 相类似的结论：三个深度特征的鲁棒性排序为：DeepID＞LCNN＞VGG，且在深度特征域中，带权分类器相对于通用分类器的性能优势不像在浅层特征域中那样明显.

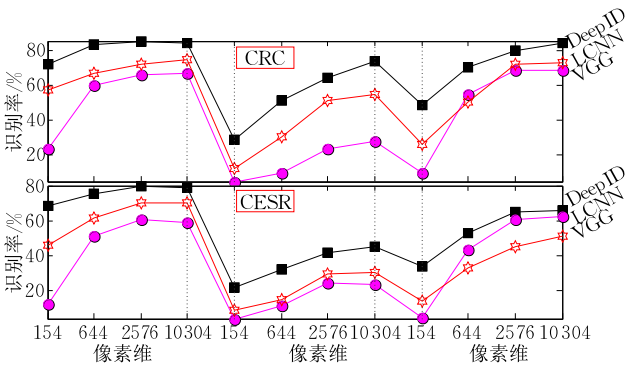


图 26 通用分类器 CRC 和带权分类器 CESR 使用三种 CNN 特征(DeepID/VGG/LCNN)在 AR 库上的识别率(从左至右依次为测试集 I、II、III)

表 5 统计了在 AR 库的三个测试集上所能达到的最优识别率及相应的分类器和特征的组合. 可以看出，没有哪个分类器与特征的组合在所有测试集上都是最优的. 擅长处理富有纹理和方向变化的遮挡的 NN＋PCANet 并不一定擅长处理缺乏纹理变化的强光遮挡，而且可能发生严重的性能退化；而擅长处理缺乏纹理变化的强光遮挡的 CRC＋Weber 和 RNR＋Weber 却在处理具有纹理变化的实体遮挡时仍有不足. 相对来说，MEC＋IGO 在各个测试集上的识别性能比较折衷，但在测试集 II 上 MEC＋IGO 与 NN＋PCANet 有较大的差距，图 24(下图)对此进行了分析，可以发现：归根结底还是因为 IGO 特征缺乏足够丰富的判别信息. 如何在有遮挡情形下缩小图像在鲁棒特征域中的类内距离并同时扩大类间距离，仍然是有待于进一步研究的问题.

表 5 AR 数据库的三个测试集上的最优识别率/%(每个识别率下方给出了达到该识别率的分类器与特征的组合)

测试集	154	644	2576	10304
I	74.79	92.72	97.20	99.44
	CRC＋Weber	RNR＋Weber	RNR＋Weber	MEC＋IGO
II	49.86	88.52	91.32	81.23
	NN＋PCANet	NN＋PCANet	NN＋PCANet	MEC＋IGO
III	69.19	87.39	96.92	95.52
	NN＋PCANet	MEC＋IGO	MEC＋IGO	MEC＋IGO

7.4 基于 LFW 数据库的非受控环境下的有遮挡人脸识别

本节验证表 2 和表 3 中的鲁棒特征和鲁棒分类器在非受控的 LFW 数据库上的识别性能. LFW 数据库[106]包含了 5749 个人的 13233 张人脸图像，所有图像均采集于互联网，被广泛应用于非受控环境下的人脸识别与认证算法的有效性验证[18,37,64-65,91,93,95,110-111]. 由于 LFW 的人脸图像包含了较大的姿势、表情、遮挡和光照等变化，图像的预处理对于算法性能的提升就特别重要. 本文采用由 Huang 等人[112]提供的基于深度学习方法对齐的 LFW 人脸图像数据集 LFW-deepfunneled，所有图像被裁剪和变换为 62×58 的灰度像素矩阵. 选取样本个数大于 8 的 217 个人用于识别实验，其中，训练集包含 217×8＝1736 张图像，测试集包含 3086 张图像，具体选取策略如图 27 所示.

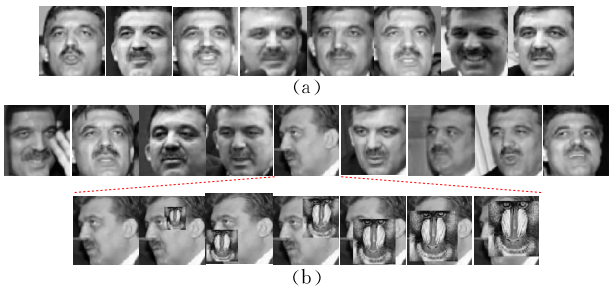
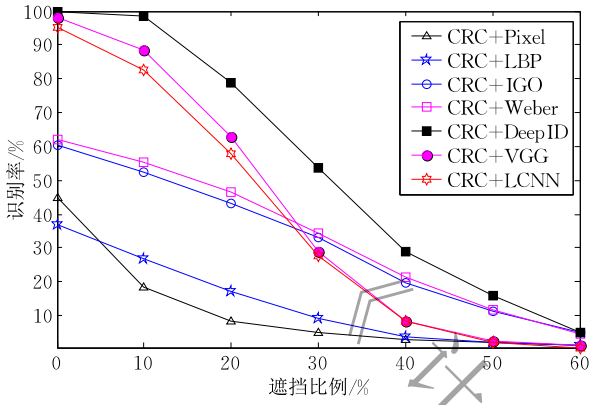


图 27 本节实验所采用的来自于 LFW 人脸库[106]的训练数据和测试数据样例((a)训练集：具有较少的光照和姿势变化且不含遮挡，每个人包含 8 个训练样本；(b)测试集：具有较大的姿势变化且包含了一定的自然遮挡和 0～60% 的模拟遮挡(第 2 行)，所有训练集之外的样本均用于测试)

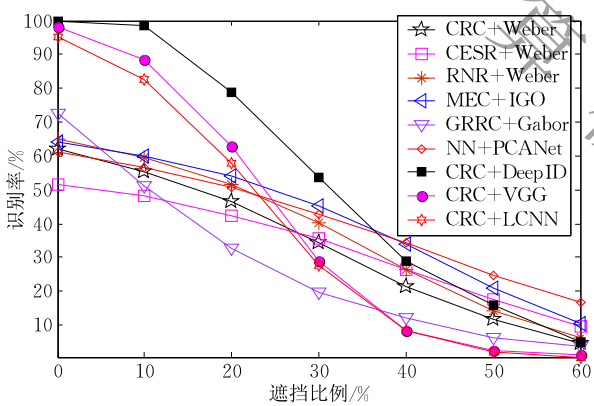
图 28(a)比较了四种浅层特征(Pixel/LBP/IGO/Weber)和三种 CNN 特征(DeepID/VGG/LCNN)在相同的分类器 CRC 下的识别性能：四种浅层特征的鲁棒性排序大体上仍然为：Weber＞IGO＞LBP＞Pixel；三种 CNN 特征的鲁棒性排序为：DeepID＞VGG＞LCNN. 在 LFW 的测试集上，三种 CNN 特征首次凸显出了其性能优势：DeepID 在各遮挡比例下的识别性能都是最优的；当遮挡比例较低(＜30%)时，VGG 和 LCNN 的识别性能要远远优于四种浅层特征；而当遮挡比例升高(＞30%)时，三种 CNN 特征的识别性能迅速衰减，IGO 和 Weber这两种浅层特征开始逐渐凸显出其性能优势.

浅层特征和深度特征在低水平遮挡和高水平遮挡下的鲁棒性差异通过图 28(a)的性能比较可以更为明显地观察到. 图 28(b)将各浅层特征和各鲁棒分类器的最优组合与四种深度特征(加入了 NN＋PCANet)相比较. 当遮挡比例较高(＞40% 或＞50%)时，MEC＋IGO 和 CESR＋Weber 的识别率逐渐超过了 CRC＋DeepID，这再次表明鲁棒特征与鲁棒分类器的结合的必要性. 另外，值得注意的是，当模拟遮挡的比例为 0 时，GRRC＋Gabor 在众多的浅层特征表现出了最优的识别性能，这是因为与

其它浅层特征所使用的滤波器相比，Gabor 滤波器是唯一模拟了 V1 视觉区域的感受野响应的滤波器[88]. 因此可以这样理解：模拟了人类视觉机制的 Gabor 滤波器（从视觉细胞感受野响应的角度[88]）和深度网络（从人类视觉对输入信号的分层次处理的角度[78]）都在一定程度上提升了非受控环境下的人脸识别性能. 但据我们所知，目前尚未有学者将两者结合起来深入究.



(a) CRC基于四种浅层特征和三种CNN特征的识别率



(b) 基于各特征与分类器的最优组合的识别率

图 28　在 LFW 上的识别率

图 28(a)和(b)的实验结果都表明，即使在非受控环境中，浅层特征中的遮挡处理机制仍然能够发挥一定的作用并值得在进一步研究中借鉴. 表 6 总结了在不同遮挡比例下的最优识别率. 比较表 4、表 5 和表 6 可以看出，在非受控环境下遮挡对识别性能的影响要远远大于受控环境，主要原因在于：非受控环境下的人脸图像受到了诸如表情、光照、姿势等更多变化因素的影响，而当这些变化因素进一步与遮挡混合时所造成的识别困难会远远超过受控环境下的只有光照和遮挡混合的情形.

表 6　LFW 上的最优识别率/%（DeepID 和 PCANet 分别取得了 0～30%和 40%～60%的遮挡比例下的最优识别率）

0	10%	20%	30%	40%	50%	60%
100	98.74	78.99	53.48	34.18	24.64	16.87

下面我们进一步从 CNN 网络的响应及其稀疏性来分析姿势变化和遮挡对网络性能的影响. 图 29 比较了三种 CNN 特征对同一个人的有姿势变化和不同程度遮挡的面部图像的神经元响应. 可以发现：低水平遮挡对卷积网络的输出影响较小，而高水平遮挡会在不同程度上扰乱网络的输出；显著的姿势变化会引发人脸图像的“自遮挡”，并且可能会比低水平的“实物遮挡”在更大程度上扰乱神经网络的输出. 我们在 5.4 节分析了网络输出的稀疏性和鲁棒性之间的关系. 从图 29 中神经元的响应来看，姿势变化（侧面脸）会使得各神经元的响应变得“稀疏”，而遮挡则会在不同程度上破坏这种稀疏性，其中以 LCNN 的稀疏性变化最为显著. Sun 等人[37]近来正致力于稀疏神经网络的研究，我们期待这一成果在对遮挡问题的处理上能够发挥更大的作用.

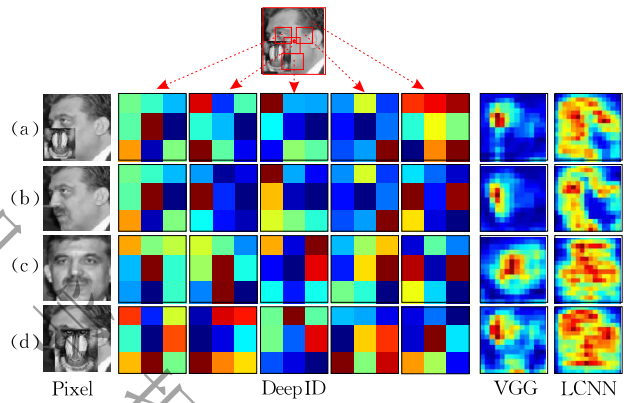


图 29　三种 CNN 特征（DeepID/VGG/LCNN）在卷积层（conv4/conv5_3/mfm5）对 LFW 中同一个人的具有姿势变化和不同程度遮挡的人脸图像的神经元响应[①]（(a)、(b)、(d) 针对包含了 20%、0、40%遮挡的侧面脸；(c) 针对不含遮挡的正面脸）

8　仍然存在的问题与进一步研究的重点

8.1　遮挡理解与人类视觉

有遮挡人脸识别仍然存在如下三个关键问题：如何判断一张人脸图像中有无遮挡？如何迅速地定位遮挡？如何消除因遮挡而引发的混叠？上述问题归根结底是一个问题：究竟什么是遮挡. 事实上，遮挡在视觉上容易判断，但在理论上却难以建模. 这是因为遮挡与具有特定形体的目标（如人脸、车辆等）不同：遮挡并没有固定的形式和内容，它只是一种外

① 为便于显示，图中采用了 Taigman 等人[92]显示神经元响应的方法：将每一个卷积层的各通道的神经元响应叠加了起来.

来的覆盖.在一般情况下,可以参照人眼的感知来"定义"遮挡,例如,对于图 30(a)和(b),人眼可以迅速做出判断:它们分别包含了"面具"和"手"的遮挡;对于图 30(c),人眼知道其所佩戴的框架眼镜虽然是一种外来的附加物,但不是遮挡;对于图 30(d),人眼认为它根本就不是一张人脸,但事实上,它只是图 30(e)中的人脸图像与一个巨大的脉冲信号的叠加而已.在上面的例子中,机器视觉与人类视觉恰好相反,机器很难准确地检测到图 30(a)和(b)中的遮挡,也很容易对图 30(c)中的框架眼镜做出误判(认为其是遮挡),但图 30(d)中的脉冲信号基本不会干扰机器的识别.如何让机器像人眼那样"理解"遮挡,是有遮挡人脸识别问题面临的终极挑战.

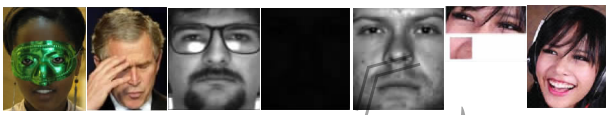


图 30 遮挡存在的几种形式与人眼对遮挡的理解

我们在前文提到现有的方法重在检测并排除遮挡,而遮挡重建方法一般效果不佳.然而,人类视觉在识别有遮挡图像时却并未将遮挡部位的信息彻底丢弃掉,而是用了一种强大的视觉推断能力部分地填补了所缺失的遮挡部位的属性信息.Luo 等人[71]曾给出了一个如图 30(f)和(g)所示的例子:人眼可以根据"眼睛-鼻子-嘴巴"这一线条的相关信息来推断该张人脸的表情属性(是否为微笑等),即使在该线条中的某一部分特征(如嘴巴)缺失的情况下.如何根据人类的视觉推理机制来推断被遮挡区域的高阶属性(如表情等)也属于遮挡理解的范畴,挑战着机器视觉的自动推理能力.

8.2 鲁棒误差编码模型中的优化问题

为了求解鲁棒误差编码模型(5)和(13),通常需要通过交替迭代估计重构系数 x 和误差权重 $\boldsymbol{\omega}$(或遮挡支撑 s).4.1 节讨论了交替迭代的两种策略:以 CESR[10]为代表的方法将误差权重 $\boldsymbol{\omega}$ 的迭代过程"嵌入"到 x 的迭代过程中;以 RSC[14]和 SSEC[9]等为代表的方法将重构 x 和重构权重 ω(或遮挡支撑 s)看作两个孤立的过程.本文第 7 节的实验表明,在鲁棒特征空间中,第一种策略通常比第二种策略更为经济、有效.因此,未来研究的一个重点是:将鲁棒的交替迭代回归模型都迁移到第一种优化策略框架下求解.

收敛性与迭代停止条件的关系.收敛性对于优化问题求解的重要性是不言而喻的,但对于识别问题而言,收敛可能意味着过拟合.在图 21 和图 22 中,我们都看到了这样的例子:算法有时并非完全不能识别,只是没有在正确识别之前及时停止.如果一旦正确识别就停止迭代,我们发现许多识别算法的性能都将得到大幅度提升.因此,如何在算法迭代过程中加入一定的监督信息,使之能够及时停止,也是未来研究的一个重要任务.

8.3 遮挡过滤与遮挡不变特征

研究者们在有遮挡人脸识别领域中的研究对于图像特征提取有两个突出的贡献:以 Tzimiropoulos 等人[33]所提出的 IGO 为代表的浅层特征提取方法及其所触发的图像差异的"零和"现象(5.2 节),开启了自动滤除异常特征的新思路;以 Chan 等人[17]提出的 PCANet 和 Sun 等人[18]提出的 DeepID 为代表的深层特征学习方法,开启了提取"遮挡不变"特征的新思路.两种思路都没有对遮挡做明确的假设或显式地处理,却都在一定程度上达到了对遮挡鲁棒的效果.这就启发我们进一步探索这两种思路背后的理论依据及其所存在的问题,例如,为什么在 IGO 域中不同图像之差会以很高的置信度服从均匀分布?为什么经过强监督学习的深度网络就可以对异常特征"免疫"?IGO 特征能够自动地将两幅图像中局部不相同的特征过滤掉,但对于局部相似而非完全不同的特征又该如何处理?PCANet 在处理具有低频噪声的图像时仍然存在缺陷,并且在高维情形下存在严重的性能退化问题.显然,所有的方法都有其局限性,在我们深刻理解其原理之前,这种潜在的局限性都是存在的.

8.4 卷积神经网络在处理遮挡问题上面临的挑战

在卷积神经网络基于大规模训练数据日益向更深和更轻的方向发展的今天,我们注意到其在处理人脸遮挡问题上仍然面临着如下挑战:

(1)将在大规模非受控环境下训练得到的深度网络模型应用于受控环境下的有遮挡人脸识别.这一问题的主要困难在于:受控环境下的训练样本通常较少,且测试数据与训练数据发生了严重的偏移.

(2)基于小样本训练异常检测网络.尽管深度学习借鉴了人类视觉皮层从低阶到高阶"认知"信号的过程[78],但就遮挡问题而言,人类视觉与深度学习背道而驰的一点是:人类视觉不需要大规模训练,人眼只需要观察几张人脸图像,就可以知道一张新的图像是否是"人脸"以及一张人脸图像中是否存在遮挡.如何训练基于小样本的异常检测网络仍然是具有很大的挑战性的问题.

(3)基于已有的误差编码方法和鲁棒特征提取方法提升深度网络模型处理遮挡的能力.传统的机器学习思想正在越来越多的被借鉴到深度学习中,如基于有监督信号的度量学习[18,65]、多尺度变换[94]、稀疏性约束[37]等,那么,是否可以将已有的处理遮挡问题的误差编码方法和传统的特征提取方法用于深度学习以提升其处理遮挡问题的能力?

(4)神经网络的稀疏性及其对遮挡的鲁棒性.鉴于人类视觉强大的信号理解能力和容错能力,在经典的计算机视觉领域中,基于手工设计的特征提取过程往往需要借鉴或者模拟人类视觉机理,其中最典型的就是 Gabor 滤波器[88].1996 年,Olshausen 等人[113]首次通过学习的方法获得了与简单细胞的感受野响应具有类似结构的字典原子.那么,经过大规模数据训练得到的神经网络的卷积核是否也有相同的性质?从公开的 VGG 和 LCNN 以及本文训练所得的 DeepID 等网络模型的卷积核结构来看,这一性质是不满足的.2016 年,Sun 等人[37]基于稀疏性约束训练 CNN 网络,并在使用同等规模的训练集的前提下取得了比稠密网络模型更好的性能.我们期待这一成果在未来对遮挡问题的处理上能够发挥更大的作用.

9　结　论

有遮挡人脸识别问题是面向现实的人脸识别系统面临一个重要挑战.本文从子空间回归、结构化误差编码、迭代重权误差编码和鲁棒特征提取等方面回顾了现有的有遮挡人脸识别方法,探讨了各类方法的产生动机、主要模型和优缺点.在三个基准数据库上对现有方法进行了大规模测试,指出了现有的鲁棒分类方法和鲁棒特征提取技术所适用的问题及其局限性,展望了进一步的研究重点,旨在吸引更多的学者关注这一领域,使有遮挡人脸识别问题在理论和实践上都得到更好的解决,并推动面向现实的人脸识别技术使其日益成熟.

参　考　文　献

[1] Zhao W, Chellappa R, Phillips P J, et al. Face recognition: A literature survey. ACM Computing Surveys, 2003, 35(4): 399-458

[2] Quionero-Candela J, Sugiyama M, Schwaighofer A, et al. Dataset Shift in Machine Learning. Cambridge, USA: The MIT Press, 2009

[3] Sun Y, Liang D, Wang X, et al. DeepID3: Face recognition with very deep neural networks. Hong Kong, China: The Chinese University of Hong Kong, Technical Report: arXiv: 1502. 00873, 2015

[4] Ding C, Xu C, Tao D. Multi-task pose-invariant face recognition. IEEE Transactions on Image Processing, 2015, 24(3): 980-993

[5] Ding C, Tao D. A comprehensive survey on pose-invariant face recognition. Sydney, Australia: University of Technology, Technical Report: arXiv:1502.04383, 2015

[6] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227

[7] Li X, Liang R, Feng Y, et al. Robust face recognition with occlusion by fusing image gradient orientations with Markov random fields//Proceedings of the International Conference on Intelligence Science and Big Data Engineering. Suzhou, China, 2015: 431-440

[8] Wei X, Li C, Hu Y. Robust face recognition with occlusions in both reference and query images//Proceedings of the International Workshop on Biometrics and Forensics. Lisboa, Portugal, 2013: 1-4

[9] Li X, Dai D, Zhang X, et al. Structured sparse error coding for face recognition with occlusion. IEEE Transactions on Image Processing, 2013, 22(5): 1889-1900

[10] He R, Zheng W, Hu B. Maximum correntropy criterion for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1561-1576

[11] Luo L, Yang J, Qian J, et al. Nuclear-L_1 norm joint regression for face reconstruction and recognition with mixed noise. Pattern Recognition, 2015, 48(12): 3811-3824

[12] Zhou Z, Wagner A, Mobahi H, et al. Face recognition with contiguous occlusion using Markov random fields//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1050-1057

[13] Yang M, Zhang L, Shiu S C, et al. Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary. Pattern Recognition, 2013, 46(7): 1865-1878

[14] Yang M, Zhang L, Yang J, et al. Robust sparse coding for face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado, USA, 2011: 625-632

[15] Zafeiriou S, Tzimiropoulos G, Petrou M, et al. Regularized kernel discriminant analysis with a robust kernel for face recognition and verification. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(3): 526-534

[16] Yang M, Zhang L, Yang J, et al. Regularized robust coding for face recognition. IEEE Transactions on Image Processing, 2013, 22(5): 1753-1766

[17] Chan T, Jia K, Gao S, et al. Pcanet: A simple deep learning baseline for image classification? IEEE Transactions on Image Processing, 2015, 24(12): 5017-5032

[18] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2892-2900

[19] Liang R, Li X. Mixed error coding for face recognition with mixed occlusions//Proceedings of the International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3657-3663

[20] Deng W, Hu J, Guo J. Extended SRC: Undersampled face recognition via intra-class variant dictionary. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(9): 1864-1870

[21] Luan X, Fang B, Liu L, et al. Extracting sparse error of robust pca for face recognition in the presence of varying illumination and occlusion. Pattern Recognition, 2014, 47(2): 495-508

[22] Ou W, You X, Tao D, et al. Robust face recognition via occlusion dictionary learning. Pattern Recognition, 2014, 47(4): 1559-1572

[23] Wright J, Yi M. Dense error correction via l^1-minimization. IEEE Transactions on Information Theory, 2010, 56(7): 3540-3560

[24] Zhang L, Yang M, Feng X. Sparse representation or collaborative representation: Which helps face recognition?//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 471-478

[25] Yang M, Zhang L. Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary//Proceedings of the European Conference on Computer Vision. Heraklion, Greece, 2010: 448-461

[26] Wei X, Li C, Hu Y. Robust face recognition under varying illumination and occlusion considering structured sparsity//Proceedings of the International Conference on Digital Image Computing Techniques and Applications. Fremantle, Australia, 2012: 1-7

[27] Jia K, Chan T H, Ma Y. Robust and practical face recognition via structured sparsity//Proceedings of the European Conference on Computer Vision. Florence, Italy, 2012: 331-344

[28] Nguyen N H, Tran T D. Robust lasso with missing and grossly corrupted observations. IEEE Transactions on Information Theory, 2013, 59(4): 2036-2058

[29] Qian J, Luo L, Yang J, et al. Robust nuclear norm regularized regression for face recognition with occlusion. Pattern Recognition, 2015, 48(10): 3145-3159

[30] Starck J, Elad M, Donoho D L. Image decomposition via the combination of sparse representations and a variational approach. IEEE Transactions on Image Processing, 2005, 14(10): 1570-1582

[31] He R, Zheng W S, Tan T, et al. Half-quadratic-based iterative minimization for robust sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(2): 261-275

[32] Yang M, Zhang L, Shiu S, et al. Robust kernel representation with statistical local features for face recognition. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(6): 900-912

[33] Tzimiropoulos G, Zafeiriou S, Pantic M. Subspace learning from image gradient orientations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(12): 2454-2466

[34] Jie C, Shiguang S, Chu H, et al. WLD: A robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1705-1720

[35] Ekenel H, Stiefelhagen R. Why is facial occlusion a challenging problem?//Proceedings of the Advances in Biometrics. Heidelberg, Germany, 2009: 299-308

[36] Martinez A M. The AR face database. Columbus, USA: Computer Visual Center, Ohio State University, Technical Report: 24, 1998

[37] Sun Y, Wang X, Tang X. Sparsifying neural network connections for face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4856-4864

[38] Masi I, Trăn A T, Hassner T, et al. Do we really need to collect millions of faces for effective face recognition?//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 579-596

[39] Wei C, Chen C, Wang Y F. Robust face recognition with structurally incoherent low-rank matrix decomposition. IEEE Transactions on Image Processing, 2014, 23(8): 3294-3307

[40] Donoho D L. Compressed sensing. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306

[41] Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing, 2006, 54(11): 4311

[42] Wright J, Ganesh A, Rao S, et al. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2009: 2080-2088

[43] Candes E J, Li X, Ma Y, et al. Robust principal component analysis? Journal of the ACM, 2011, 58(31): 1-73

[44] He R, Hu B G, Zheng W S, et al. Robust principal component analysis based on maximum correntropy criterion. IEEE Transactions on Image Processing, 2011, 20(6): 1485-1494

[45] De La Torre F, Black M J. A framework for robust subspace learning. International Journal of Computer Vision, 2003, 54(1-3): 117-142

[46] Lu C, Feng J, Chen Y, et al. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5249-5257

[47] Elhamifar E, Vidal R. Robust classification using structured sparse representation//Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition. Providence, USA, 2011: 1873-1879

[48] Kim W, Suh S, Hwang W, et al. SVD face: Illumination-invariant face representation. IEEE Signal Processing Letters, 2014, 21(11): 1336-1340

[49] Liu W, Pokharel P P, Principe J C. Correntropy: Properties and applications in non-Gaussian signal processing. IEEE Transactions on Signal Processing, 2007, 55(11): 5286-5298

[50] Jacobs D W, Weinshall D, Gdalyahu Y. Classification with nonmetric distances: Image retrieval and class representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(6): 583-600

[51] Tan X, Chen S, Zhou Z H, et al. Face recognition under occlusions and variant expressions with partial similarity. IEEE Transactions on Information Forensics and Security, 2009, 4(2): 217-230

[52] Naseem I, Togneri R, Bennamoun M. Robust regression for face recognition. Pattern Recognition, 2012, 45(1): 104-118

[53] Holland P W, Welsch R E. Robust regression using iteratively reweighted least-squares. Communications in Statistics-Theory and Methods, 1977, 6(9): 813-827

[54] Boyd S, Vandenberghe L. Convex Optimization. New York, USA: Cambridge University Press, 2004

[55] Fransens R, Strecha C, Van Gool L. Robust estimation in the presence of spatially coherent outliers//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop. New York, USA, 2006: 102

[56] Dahua L, Xiaoou T. Quality-driven face occlusion detection and recovery//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-7

[57] Yang M, Feng Z, Shiu S C K, et al. Fast and robust face recognition via coding residual map learning based adaptive masking. Pattern Recognition, 2014, 47(2): 535-543

[58] Li S, Gong D, Yuan Y. Face recognition using Weber local descriptors. Neurocomputing, 2013, 122: 272-283

[59] Yang M, Zhang L, Shiu S C, et al. Monogenic binary coding: An efficient local feature extraction approach to face recognition. IEEE Transactions on Information Forensics and Security, 2012, 7(6): 1738-1751

[60] Nguyen H, Caplier A. Local patterns of gradients for face recognition. IEEE Transactions on Information Forensics and Security, 2015, 10(8): 1739-1751

[61] Zhang B, Gao Y, Zhao S, et al. Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor. IEEE Transactions on Image Processing, 2010, 19(2): 533-544

[62] Vu N, Caplier A. Enhanced patterns of oriented edge magnitudes for face recognition and image matching. IEEE Transactions on Image Processing, 2012, 21(3): 1352-1365

[63] Vu N. Exploring patterns of gradient orientations and magnitudes for face recognition. IEEE Transactions on Information Forensics and Security, 2013, 8(2): 295-304

[64] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2014: 1988-1996

[65] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1891-1898

[66] Liu C, 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

[67] Ren C, Dai D, Li X, et al. Band-reweighed Gabor kernel embedding for face image representation and recognition. IEEE Transactions on Image Processing, 2014, 23(2): 725-740

[68] Zhang B, Shan S, Chen X, et al. Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition. IEEE Transactions on Image Processing, 2007, 16(1): 57-68

[69] Zhang W, Shan S, Gao W, et al. Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition//Proceedings of the IEEE International Conference on Computer Vision. Beijing, China, 2005: 786-791

[70] Kumar N, Berg A C, Belhumeur P N, et al. Attribute and simile classifiers for face verification//Proceedings of the International Conference on Computer Vision. Kyoto, Japan, 2009: 365-372

[71] Luo P, Wang X, Tang X. A deep sum-product architecture for robust facial attributes analysis//Proceedings of the IEEE International Conference on Computer Vision. Portland, USA, 2013: 2864-2871

[72] Kumar N, Berg A, Belhumeur P N, et al. Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(10): 1962-1977

[73] Zhang N, Paluri M, Ranzato M, et al. PANDA: Pose aligned networks for deep attribute modeling//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1637-1644

[74] Ungerleider L G, Haxby J V. 'what' and 'where' in the human brain. Current Opinion in Neurobiology, 1994, 4(2): 157-165

[75] Felleman D J, Van Essen D C. Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex, 1991, 1(1): 1-47

[76] Turk M, Pentland A. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86

[77] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788-791

[78] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[79] Li S Z, Xin W H, Hong J Z, et al. Learning spatially localized, parts-based representation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Kauai, USA, 2001: 207-212

[80] Zhang T, Tang Y Y, Fang B, et al. Face recognition under varying illumination using Gradientfaces. IEEE Transactions on Image Processing, 2009, 18(11): 2599-2606

[81] Wang B, Li W, Yang W, et al. Illumination normalization based on Weber's law with application to face recognition. IEEE Signal Processing Letters, 2011, 18(8): 462-465

[82] Gross R, Brajovic V. An image preprocessing algorithm for illumination invariant face recognition//Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication. Guildford, UK, 2003: 10-18

[83] Savvides M, Kumar B. Illumination normalization using logarithm transforms for face authentication//Proceedings of the Audio-and Video-Based Biometric Person Authentication. Guildford, UK, 2003: 1055

[84] Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12): 2037-2041

[85] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987

[86] Ojansivu V, Heikkil A J. Blur insensitive texture classification using local phase quantization//Proceedings of the International Conference on Image and Signal Processing. California, USA, 2008: 236-243

[87] Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary patterns//Proceedings of the European Conference on Computer Vision. Prague, Czech Republic, 2004: 469-481

[88] Liu C. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(5): 725-737

[89] Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1778-1785

[90] Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 951-958

[91] Taigman Y, Yang M, Ranzato M, et al. Deepface: Closing the gap to human-level performance in face verification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1701-1708

[92] Taigman Y, Yang M, Ranzato M, et al. Web-scale training for face identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2746-2754

[93] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 815-823

[94] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[95] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition//Proceedings of the British Machine Vision Conference. Swansea, UK, 2015: 41.1—41.12

[96] Wu X, He R, Sun Z. A lightened CNN for deep face representation. Beijing, China: University of Science and Technology Beijing, Technical Report: arXiv:1511.02683, 2015

[97] Lecun Y, Bottou L E O, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[98] Sun Y, Wang X, Tang X. Hybrid deep learning for face verification//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 1489-1496

[99] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-14

[100] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[101] Gao S, Tsang I W, Chia L. Sparse representation with kernels. IEEE Transactions on Image Processing, 2013, 22(2): 423-434

[102] Wang D, Lu H, Yang M. Kernel collaborative face recognition. Pattern Recognition, 2015, 48(10): 3025-3037

[103] Naseem I, Togneri R, Bennamoun M. Linear regression for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(11): 2106-2112

[104] Jia Y, Shelhamer E, Donahue J, et al. Caffe: convolutional architecture for fast feature embedding//Proceedings of the ACM International Conference on Multimedia. Orlando, USA, 2014: 675-678

[105] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild//Proceedings of the International Conference on Computer Vision. Santiago, Chile, 2015: 3730-3738

[106] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Amherst, USA: University of Massachusetts, Technical Report: 07-49, 2007

[107] Ghazi M M, Ekenel H K. A comprehensive analysis of deep learning based representation for face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA, 2016: 34-41

[108] Lee K C, Ho J, Kriegman D J. Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 684-698

[109] Georghiades A S, Belhumeur P N, Kriegman D J. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 643-660

[110] Pinto N, Dicarlo J J, Cox D D. How far can you get with a modern face recognition test set using only simple features?// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 2591-2598

[111] Guillaumin M, Verbeek J, Schmid C. Is that you? Metric learning approaches for face identification//Proceedings of the International Conference on Computer Vision. Kyoto, Japan, 2009: 498-505

[112] Huang G B, Mattar M, Lee H, et al. Learning to align from scratch//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 773-781

[113] Olshausen B A. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 1996, 381(6583): 607-609

LI Xiao-Xin, born in 1980, Ph. D., associate professor. His research interests include computer vision, pattern recognition, and image processing.

LIANG Rong-Hua, born in 1974, Ph. D., professor. His research interests include computer vision and image processing.

Background

Face recognition is one of the most challenging problems in computer vision and pattern recognition. Practical face recognition system has to struggle with many variations, such as illumination, pose and partial occlusions. This work mainly focuses on face recognition with partial occlusions.

The main difficulties incurred by occlusion include feature missing, alignment error and local alias. One main classical solving scheme is to build error coding models to explicitly represent, weaken or eliminate the errors incurred by occlusion, such as RNR, CESR, and SSEC. These methods are originally designed in pixel domain but have limited generalization performance. Very recently, researchers began to pay attention to design occlusion-robust features. The well-known occlusion-robust features include both the shallow ones, such as IGO and Weberfaces, and the deep ones, such as PCANet and DeepID. One of the amazing results claimed by the authors is that even the methods used to extract these features are not taught to distinguish occlusions from faces, they could either automatically eliminate occlusion, such as IGO and Weberfaces, or produce occlusion-invariant features, such as PCANet and DeepID. However, these robust feature extraction methods are not the final solution. First, the principles lying behind these feature extractions are still not very clear. Second, experiments show that the occlusion-invariant properties are not always true especially when the occlusion is caused by highlight illuminations. Third, the efficacy of these feature extraction methods is only verified in experiment but not proved in theory.

We therefore deeply study the extant methods in detail. The main contributions of this review are in 3 aspects. First, we summarize the extant methods into several distinct clues, along which the readers could easily find their interesting points. Second, we build many connections which are not very obvious before, such as sparse representation and image decomposition, illumination-invariant features and occlusion-robust ones, the deep learning network, PCANet, and the classical shallow feature extraction framework (Feature-Pattern-Histogram). Third, we experimentally show how robust features help robust classifiers improve their recognition performance and also analyze how the extant classifiers fail to utilize the robust features. We expect that the above contributions could inspire the readers to find new interesting research topic in the future.

This work is partially supported by the National Natural Science Foundation of China (61402411, 61379017 and 61672464), and the Zhejiang Provincial Natural Science Foundation (LY18F020031, LY17F020021). Face recognition with occlusion is a main research topic of our research group in the Zhejiang University of Technology. In the past few years, we have published a series of papers directly related with this research topic in *IEEE Transactions on Image Processing*, International Joint Conferences on Artificial Intelligence and etc. And we have also acquired the warranty of a related Chinese patent.