基于深度特征的快速协同表示人脸识别

陈超, 宋磊

南京大学工程管理学院 *2017* 级，南京 *210046*

摘要　在人脸识别中，人脸图像受光线，表情，遮挡，姿态变化影响. 此外，现实应用中少量的训练样本也给人脸识别增加了难度. 为提高人脸识别的准确率，本文提出了将协同表示分类器应用在深度学习特征的人脸识别算法. 该算法首先利用已训练好的深度卷积神经网络提取人脸特征，再采用一种新的特征选择算法对特征分量进行筛选，建立不同人脸的特征字典. 最后，在求解测试样本特征的协同表示系数的基础上，根据范数进行分类. 实验在 FERET,Extended Yale B,WebFace 数据集上进行测试，结果证明新提出的算法有很好的分类效果.

关键词　深度学习；特征提取；人脸识别；特征选择；协同表示

引言

人脸识别是近年来计算机视觉和模式识别领域的一个研究热点[1]. 由于光照，表情，遮挡，噪声，姿态变化等存在[2]，人脸识别过程中存在着诸多挑战. 小样本问题也给人脸识别带来了困难. 因此如何达到良好且稳定的人脸识别效果目前依然是研究人员关注的重点.

随着研究的深入和发展，学者们提出了许多经典的人脸识别算法. 基于稀疏表示的理论，Wright 等人[3] 利用 ℓ_1 范数最小化求最稀疏的解进行人脸识别. 由于 ℓ_1 范数优化耗时长，而且学者对稀疏性是否有利于解决小样本问题存在着疑问[4]，因此一种基于更加松弛的 ℓ_2 范数协同表示分类器 (Collabrative representation classifier, CRC) 被提出[4]，在不充分的训练字典问题中取得了一定突破.

为了充分挖掘人脸图像的信息，常常需要对人脸特征进行提取. 目前，常用的人工特征提取方式包括 Gabor 小波变换[5]、局部二值模式等. 人工特征表达能力有限，而且分类精度受人为因素的干扰，因此在实际应用中受到限制. 随着深度学习的发展，人们开始关注深度卷积神经网络的特征提取能力[6, 7]. 深度网络能够将图像映射到一个特征子空间内，映射后的特征具有良好的可分性和鲁棒性. 研究表明，深度特征有利于提高分类问题的准确率[12].

基于以上，本文提出了一种融合深度学习和协同表示的快速鲁棒人脸识别算法. 该算法复杂的场景下的小样本人脸识别问题中仍然能够保持较高的识别准确率. 本文主要贡献如下：

1. 利用协同表示处理深度学习特征，充分发挥了深度学习在特征提取上的优点和协同表示在小样本识别上的优点.
2. 利用范数代替残差进行分类，在保证准确率相近的情况下大幅提升识别速度
3. 提出一种与协同表示相适应的特征选择算法，能够筛选出有效的特征从而降维.

1　相关工作

1.1　深度卷积神经网络

近年来，随着计算机能力的提高，深度卷积神经网络 (deep convolutional neural network, DCNN) 成为图像识别领域的热门方向. DCNN 拥有强大的函数表达能力. Hinton 等[8] 使用大规模深度卷积网络在 Imagenet 这样 1000 类的分类问题上取得了非常好的结果. 文献[9, 10, 11] 提出了不同的网络结构，均取得了显著的效果，进一步推动了图像识别的发展.

DCNN 每个层都包含不同层次的特征. 较浅层主要包含图像的纹理和边角等局部特征，较深层包含的是更能代表类别的全局特征. 随着层的逐渐加深，特征越来越复杂. 经典的网络如 Alexnet[8] 在最后几层会使用全连接层融合提取到的深度特征.

在人脸识别领域，DCNN 虽然拥有令人瞩目的表现，但它需要大量训练样本，而且训练时间也相对较长. 实际生产应用中一个人的训练样本往往有限，这使得 DCNN 的应用受到限制.

1.2 稀疏表示人脸识别

稀疏表示实质就是将待测信号 y 用字典 X 线性表示，即 $y \approx X\alpha$，其中 α 为稀疏系数矢量. α 的稀疏性由 ℓ_0 来衡量. 由于 ell_0 范数优化为 NP-hard 问题，常用 ℓ_1 范数或 ell_2 范数来代替 ℓ_0 范数. 尽管 ℓ_1 范数更接近 ℓ_0 范数，能够获得更稀疏的系数，但依然存在这计算复杂度较高的问题. 在实时人脸识别系统中，基于 ℓ_2 范数的稀疏表示 (即协同表示) 更受欢迎.

基于稀疏表示的人脸识别方法假设同一类人脸位于同一子空间. 这一假设在较复杂的环境中（如姿态，表情，光线，遮挡等）一般不成立. 为了减少干扰的影响，文献[13, 14, 15, 17] 采取了一系列措施，在特定的问题上取得了一定突破，但在大规模的无限制人脸识别应用中表现依然劣于深度学习方法.

2 基于深度学习和稀疏表示的快速鲁棒人脸识别

本文提出了兼顾二者所长的基于深度学习的特征和快速协同表示的人脸方法（fast collaborative representation via deep feature, FCRDF）. 相较于传统的稀疏表示算法，FCRDF 对光线、姿态变化等干扰更鲁棒. 相较于深度学习方法，FCRDF 不需要大量的训练数据，也不需要长时间地训练或微调网络. FCRDF 整体流程包括特征提取、特征选择和分类三个环节.

2.1 特征提取

FCRDF 在特征提取阶段采用基于 vgg-16 的网络结构并使用开源的已训练好的权重文件[16]. 该网络的训练集来源于谷歌图片. 网络各层具体信息如表 1所示. 原始图像经过该网络得到 2622 维的深度

网络层	说明	参数
Input	输入层	224× 224 大小的人脸 RGB 格式图像
Conv1,2	卷积层	64 个 3×3 大小的卷积核，步长为 1
MP1	最大池化层	2×2 大小的范围，步长为 2
Conv3,4	卷积层	128 个 3×3 大小的卷积核，步长为 1
MP2	最大池化层	2×2 大小的范围，步长为 2
Conv5,6,7	卷积层	256 个 3×3 大小的卷积核，步长为 1
MP3	最大池化层	2×2 大小的范围，步长为 2
Conv8,9,10	卷积层	512 个 3×3 大小的卷积核，步长为 1
MP4	最大池化层	2×2 大小的范围，步长为 2
Conv11,12,13	卷积层	512 个 3×3 大小的卷积核，步长为 1
MP5	最大池化层	2×2 大小的范围，步长为 2
FC1	全连接层	输入 25088 维，输出 4096 维
FC2	全连接层	输出 4096 维
FC3	全连接层	输出 2622 维

Table 1: 特征提取网络结构

特征向量.

Algorithm 1 F eatureSelectForCRC

输入: 原始人脸特征集 $\mathbf{F}=\{\mathbf{f_1},\mathbf{f_2},\cdots,\mathbf{f_m}\}$, 重复次数 η

输出: 相关统计量 $\boldsymbol{\theta}=\{\theta_1,\theta_2,\cdots,\theta_n\}$

根据类别将 $\mathbf{F}$ 分成 c 个特征子集: $\mathbf{F_1},\mathbf{F_2}\cdots,\mathbf{F_c}$, $|\mathbf{F_i}|=\frac{m}{c}$
初始化 $\boldsymbol{\theta}=\{0,0,\cdots,0\}$
for $i=1,2,\cdots,c$ **do**
    for $j=1,2,\cdots,\eta$ **do**
        将 $\mathbf{F_i}$ 随机划分成数量相等的两份 $\mathbf{Ftrain_i}$ 和 $\mathbf{Ftest_i}$
        for $k=1,2,\cdots,\frac{m}{2c}$ **do**
            $\mathbf{x}=\mathbf{Ftest_{ik}}$
            $\mathbf{error}=RESIDUALOFCRC(\mathbf{x},\mathbf{Ftrain_i})$
            for $l=1,2,\cdots,n$ **do**
                $\theta_l=\theta_l+error_l$
for $i=1,2,\cdots,n$ **do**
    $\theta_i=\theta_i\times\frac{2}{m\eta}$

function $RESIDUALOFCRC(\mathbf{y},\mathbf{X})$
    $\boldsymbol{\alpha}=(\mathbf{X^T}\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X^T}\mathbf{y}$
    $\mathbf{y}^*=\mathbf{X}\boldsymbol{\alpha}$
    for $i=1,2,\cdots,n$ **do**
        $error_i=(y_i-y_i^*)^2$
    return error

2.2 特征选择

因为网络结构和训练方式等原因，提取到的特征向量里并非所有分量都对识别有贡献. 因此需要进行特征选择. 特征选择的主要目的是降维. 经典的降维方法包括主成分分析、线性判别分析[18] 等. 在小样本问题中这些方法取得的效果往往不能令人满意.

本文提出了一种适用于协同表示的特征选择算法. 设特征向量 $\mathbf{f}=\{f_1,f_2,\cdots,f_n\}$, 设置相关统计量 $\boldsymbol{\theta}=\{\theta_1,\theta_2,\cdots,\theta_n\}$, θ_i 表示运用协同表示算法后重构样本与原样本的在分量 f_i 的期望距离. 该特征选择算法描述如算法 1.

本文中，首先从 LFW 人脸数据集选出 99 个拥有 10 张图像的人，每个人随机选出 5 张图像组成训练集，另外 5 张图像组成测试集. 接下来将训练集和测试集输入网络，得到特征字典集和测试特征. 之后使用算法 1, 并且设置重复次数 $\eta=5$. 最后获得的各分量相关统计量如图所示. 小的值表示该分量能被同类样本很好地表示，因此具有较丰富的类别信息; 相反，值较大的分量可能不含类别信息或者是对噪声敏感的. 过滤掉相关统计量较大的分量，不仅可以降维，加速分类过程，还能一定程度上提高识别准确率.

算法 1只需执行一次并保存结果，之后可直接运用已有的相关统计量进行特征筛选，因此其运算开销不计入实际识别时间.

2.3 分类

对于经过特征筛选的字典集 X 和测试样本 y, 协同表示的问题形式如下:

$$\min_{\alpha} \quad \|y-X\alpha\|_2^2+\lambda\|\alpha\|_2^2$$

该问题有解析解:

$$\alpha^* = (X^T X + \lambda I)^{-1} X^T y$$

传统的协同表示算法使用残差来判断类别:

$$identity(y) = \arg\min_c \frac{\|y - X\sigma_c(\alpha^*)\|_2}{\|\sigma_c(\alpha^*)\|_2}$$

其中 $\sigma_c(\alpha^*)$ 表示在第 c 类的表示系数. 求残差涉及到矩阵运算，计算复杂度高. 为了提高识别效率，本文使用范数替代残差作为分类的依据:

$$identity(y) = \arg\max_c \|\sigma_c(\alpha^*)\|_p$$

其中 $\|\cdot\|_p$ 表示 p 范数. 类系数反映了该类图像与测试样本之间的关联程度. 同类字典集对表示测试样本的贡献更大，系数更大，因此对应的范数更大. 下面给出三种可供选择的范数:

1. ℓ_2 范数: $\|\alpha\|_2 = \sqrt{\sum_i \alpha_i^2}$
2. ℓ_1 范数: $\|\alpha\|_1 = \sum_i |\alpha_i|$
3. γ 范数: $\|\alpha\|_\gamma = \sum_i (|\alpha_i| - \frac{\alpha_i^2}{2\gamma}) I(|\alpha_i| < \gamma) + \frac{\gamma}{2} I(|\alpha_i| \geq \gamma)$

其中 $I(\cdot)$ 是指示函数. ℓ_2 范数,ℓ_1 范数和 γ 范数分别是对 ℓ_0 范数不同程度的逼近. 图 1展示了上述几种范数之间的关系.

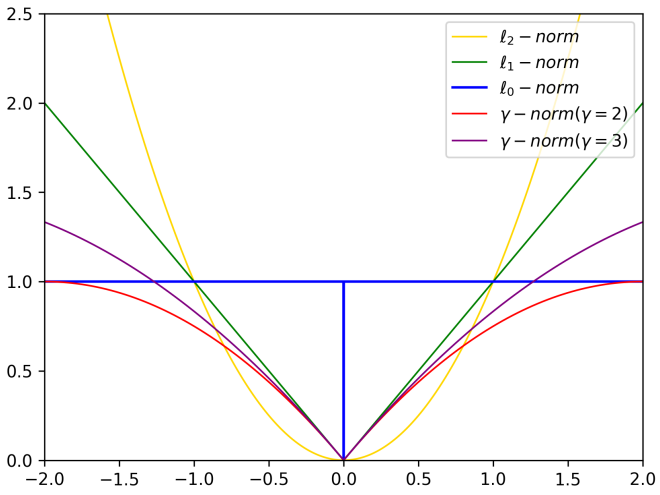


Figure 1: ℓ_2 范数、ℓ_1 范数、ℓ_0 范数和 γ 范数的对比.

为了对比范数分类和残差分类的精度和时间. 从 LFW[22] 人脸数据集选出 99 个拥有 10 张图像的人，每人选 5 张组成训练集，5 张组成测试集，分别使用残差、ℓ_1 范数、ℓ_2 范数、$\gamma(\gamma = 3)$ 分类，测量准确率和分类时间. 表 2显示了这四种方式在 LFW 数据集上的准确率和分类阶段的运行时间. 经测试处理单张图像时前面流程所花费的平均时间为 29ms，若用残差分类平均还需要 36.5ms 的时间，若用范数分类则最少还需要 2ms. 范数分类的准确率低于残差分类的 1% 左右. 可以看到范数分类能够在运行速度大幅提升的情况下保持较高的识别精度，符合实时人脸识别系统的需求.

方法 \ 指标	准确率 (%)	时间 (ms)
residual	93.54	36.56
ℓ_1 *norm*	92.32	1.94
ℓ_2 *norm*	92.12	2.41
γ *norm*	92.32	3.90

Table 2: 四种分类方式在 LFW 数据集上的准确率和运行时间

2.4 整体流程

FCRDF 整体流程可描述如下：

(1) 对训练集样本 $X = \{x_1, x_2, \cdots, x_m\}$, 通过网络映射 $f_i = f(x_i)$，获得特征空间的训练集字典 $F = \{f_1, f_2, \cdots, f_m\}$；
(2) 根据相关统计量剔除某些分量, 得到新的训练集字典 $\hat{F} = \{\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_m\}$；
(3) 对测试样本 y, 同样使用特征提取和特征选择，获得测试样本特征 $\hat{y}$；
(4) 将 $\hat{F}$ 和 $\hat{y}$ 归一化；
(5) 使用协同表示分类器，对 $\hat{y}$ 用训练集字典 $\hat{F}$ 进行协同表示，得到系数向量 $\hat{\alpha}$:

$$\hat{\alpha}^* = (\hat{F}^T\hat{F} + \lambda I)^{-1}\hat{F}^T\hat{y}$$

(6) 根据类系数向量的 p 范数进行分类：

$$identity(\hat{y}) = \arg\min_c \|\sigma_c(\hat{\alpha}^*)\|_p$$

3 实验结果及分析

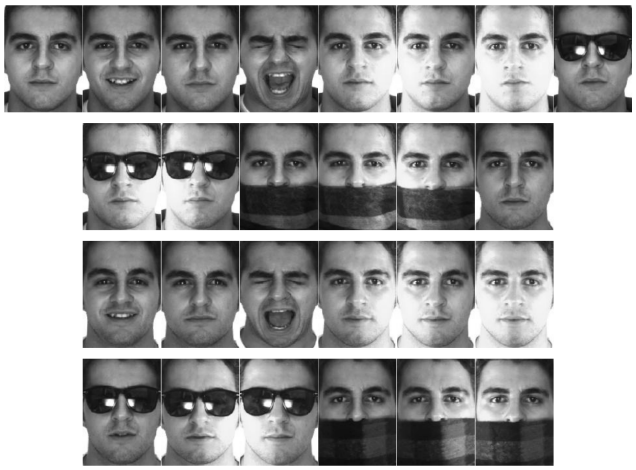
3.1 实验平台

本文使用的实验操作系统为 64 位 Linux，应用软件使用 Python2.7，Matlab2015a. 硬件平台主机 CPU 为 Inter(R) Core(TM)i5-3470 3.20GHz, 内存 8GB，此外为了加速卷积神经网络提取特征的过程，使用显存为 4G 的 NVIDIA GTX1050 显卡.

3.2 实验结果

本文分别在 FERET[19]、AR[21]、CASIA-WebFace[20] 三个人脸数据库上验证 FCRDF 的效果，同时选用对比方法为原始图像 + 最近邻分类 (Pixel+NN)、原始图像 + 协同表示 (Pixel+CRC)、深度学习特征 + 最近邻分类 (DL+NN). 为了统一，将所有图像灰度化，同时令参数 $\lambda = 4$.

(a)



(b)

(c)

Figure 2: (a)FERET 数据集 (b)AR 数据集, (c)CASIA-WebFace 数据集中的样本灰度图像

3.2.1 姿态变化下的人脸识别

选择 FERET 人脸数据库验证姿态变化下的人脸识别效果. FERET 人脸数据库由 200 个人在不同姿势下的 1400 张图像组成. 图 2(a)展示了其中一个人的所有图像. 由每个人的第一张图像组成训练集, 其余图像的组成测试集. 表 3展示了四种算法在 FERET 上的表现. 可以看出, 由于姿态变化导致子空间特性被破坏, 协同表示算法在原始图像上表现很差, 甚至不如最近邻分类. 然而提取了深度特征后, 准确率上升了 71.34%, 说明了深度学习特征的鲁棒性.

算法	准确率 (%)
Pixel+NN	41.42
Pixel+CRC	25.33
DL+NN	90.92
our method	96.67

Table 3: FERET 数据集的识别率

3.2.2 遮挡、表情、光照变化下的人脸识别

选择 AR 人脸数据集验证遮挡、表情、光照变化下的人脸识别效果. AR 人脸数据集包含 100 人的 2600 张正面近照图像. 每个人包括不同表情、光照和遮挡的人脸的正面图像. 图 2(b)展示了其中

一个人的图像. 我们分别从每个人中选取前 6 张干净的人脸图像组成容量为 228 的训练集，剩余的组成测试集. 表 4展示了四种算法在 AR 数据集上的表现. 在这种条件下协同表示的优越性显露无遗，即在较多脸部特征丢失的情况下仍然拥有不错的表现. 基于原始图像的协同表示准确率甚至高于基于深度特征的最近邻方法. 本文算法取得了 80.75% 的最佳成绩.

算法	准确率 (%)
Pixel+NN	36.80
Pixel+CRC	73.25
DL+NN	72.60
our method	80.75

Table 4: AR 数据集的识别率

3.2.3 非限制场景下的人脸识别

选择 CASIA-W ebFace 人脸数据集测试 FCRDF 算法在非限制场景下的鲁棒性. CASIA-W ebFace 人脸数据集包含 10575 人的 494414 张图像. 因为图像主要来源于网络，所以数据集组成成分十分复杂，每个人包含不同环境、不同年龄、不同清晰度的照片. 我们从中挑选了 99 人合计 7920 张图像，每个人有 80 张照片. 图像使用 DLLIB 开源库进行裁剪，并缩放到 224×224 大小. 图 2(c)展示了经过裁剪后的其中一个人的一些样本灰度图像. 我们抽取每人 10 张图像组成大小为 990 的训练集，再让其他图像当作测试图像.

算法	准确率 (%)
Pixel+NN	5.95
Pixel+CRC	11.14
DL+NN	68.17
our method	84.56

Table 5: CASIA-W ebFace 数据集的识别率

表 5展示了四种算法在 CASIA-W ebFace 数据集上的表现. 不难发现，本文提出的算法拥有最高的识别精度. 这一部分要归功于深度学习强大的特征提取能力，一部分要归功于协同表示算法对信号噪声的鲁棒性.

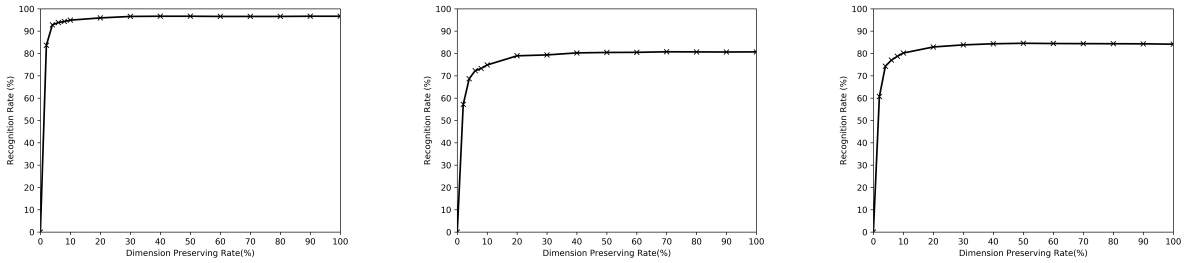
3.2.4 特征选择

为了测试算法 1的效果，分别在上述三个数据集测量不同维度下的准确率. 实验结果如图 3所示.

从图 3可以看到，利用本文提出的特征选择算法将特征维数降为原来的 20%, 各数据集的识别率几乎没有下降，即使降到原来的 10%, 识别率与原来的相比仅相差 5% 左右. 这说明原来的深度学习特征有许多冗余的分量，而算法 1能有效地将这些分量筛选出来，从而能够实现大幅度的降维.

4 结束语

本文提出了一种融合深度学习和协同表示的快速鲁棒人脸识别算法 FCRDF，将深度学习特征与协同表示结合起来，充分发挥两者优点，实现识别率的提升. 为了提高识别效率，首先设计了一种与协同表示相适应的特征选择算法，能够在维数剧烈缩减的时候维持较高的正确率. 然后在分类阶段用



(a) FERET (b) AR (c) CASIA-W ebFace

Figure 3: 不同维度下的准确率

范数代替了残差，在识别率几乎不变的情况下大大提升分类速度. 未来的工作将尝试改进网络结构，在保证特征提取能力的情况下使之轻量化，并做一些理论上的研究.

References

[1] Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: A literature survey. Acm Computing Surveys (CSUR) 35(4) (2003) 399–458

[2] Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on. Volume 1., IEEE (2005) 947–954

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. IEEE PAMI, 31(2):210–227, 2009.

[4] R. Rigamonti, M. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?" in IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1545–1552.

[5] Liu C, Wechsler H. A Gabor feature classifier for face recognition[C]//Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. IEEE, 2001, 2: 270-275.

[6] Xiong L, Karlekar J, Zhao J, et al. A good practice towards top performance of face recognition: Transferred deep feature fusion[J]. arXiv preprint arXiv:1704.00438, 2017.

[7] Xiao T, Li H, Ouyang W, et al. Learning deep feature representations with domain guided dropout for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1249-1258.

[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[9] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.

[10] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[11] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[12] Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition[C]//International conference on machine learning. 2014: 647-655.

[13] Lu C Y, Min H, Gui J, et al. Face recognition via weighted sparse representation[J]. Journal of Visual Communication and Image Representation, 2013, 24(2): 111-116.

[14] Yang M, Zhang L, Yang J, et al. Regularized robust coding for face recognition[J]. IEEE transactions on image processing, 2012, 22(5): 1753-1766.

[15] Zhang L, Zhou W D, Chang P C, et al. Kernel sparse representation-based classifier[J]. IEEE Transactions on Signal Processing, 2011, 60(4): 1684-1695.

[16] Parkhi, Omkar M. , A. Vedaldi , and A. Zisserman . "Deep Face Recognition." British Machine Vision Conference 2015 2015.

[17] Huang K, Aviyente S. Sparse representation for signal classification[C]//Advances in neural information processing systems. 2007: 609-616.

[18] Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annals of eugenics, 1936, 7(2): 179-188.

[19] Phillips P J, Wechsler H, Huang J, et al. The FERET database and evaluation procedure for face-recognition algorithms[J]. Image and vision computing, 1998, 16(5): 295-306.

[20] Yi D, Lei Z, Liao S, et al. Learning face representation from scratch[J]. arXiv preprint arXiv:1411.7923, 2014.

[21] A. M. Martinez and R. Benavente, "The AR face database," CVC, Barcelona, Spain, Tech. Rep., 1998.

[22] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]. 2008.

Fast Collaborative Representation Via Deep Feature For Face Recognition

Chen Chao, Song Lei

2017, School of Engineering, Nanjing University, Nanjing 210046

Abstract: Face recognition has a very important application in production and life. Changes in expression, light and occlusion, as well as complex environmental backgrounds, increase the difficulty of face recognition. In order to improve the accuracy of face recognition, a robust and fast face recognition algorithm combining deep learning and collaborative representation is proposed. This algorithm firstly uses the trained deep convolutional neural network to extract face features, and then uses a new feature selection algorithm to filter feature components, so as to establish feature dictionaries of different faces. The representation coefficients of the test samples are obtained by using the cooperative representation classifier, and finally the coefficients are classified according to the norm of the coefficients. Experiments are conducted on the data sets of FERET,ExtendYale B and WebFace, and the results proved that the proposed algorithm has a surprising performance.

Key words: deep learning; feature extraction; face recognition; feature selection; collabrative representation