深度学习在人脸分析与识别中的应用

山世光 阚美娜 李绍欣 等
中国科学院计算技术研究所

关键词：深度学习 人脸识别

引言

对人脸识别等视觉任务而言，特征表示和模式分类是两个核心的步骤，其中又以特征提取最为关键。过去几十年，人脸识别的发展史在很大程度上是用来表示人脸特征方法的变迁史。最早的人脸识别文献大多采用直觉上“有效”的面部几何特征描述（如嘴巴大小等）来表示人脸，但实践很快表明了其区分力的不足。1991年之后，以“Eigenfaces”为代表的子空间分析方法在人脸识别领域几乎一统天下，衍生出Fisherfaces, Laplacianfaces以及2D PCA, 2D LDA等不计其数的子空间分析方法。这些方法直接采用人脸图像中所有像素的颜色或亮度值作为初始特征，然后对它们进行“变换”得到更具区分力的人脸表示。所采用的“变换”通常是在训练集合上学习而来的，最经典的优化目标是最大化Fisher判别准则，即类内差异小且类间差异大。为克服上述方法直接以“颜色或亮度值”作为原始特征的局限性，2000年之后，涌现出很多对邻域像素亮度或颜色值进行局部特征提取的方法，其中既包括在图像分类领域取得极大成功的SIFT[1], HOG[2]等局部特征，也包括尤其适用于人脸分析的LBP[3]和Gabor特征。这类方法的共同特点是汇聚局部邻域像素团的亮度值形成局部特征，再采用子空间分析等方法对局部特征进行特征变换。在深度学习浪潮“爆发”之前，这类方法在人脸识别领域最真实的3个数据库——FERET, FRGC v2.0和LFW上均取得最高的性能。例如，在美国国家标准与技术研究院(National Institute of Standards and Technology, NIST)发布的FERET和FRGC v2.0上，性能最优的方法包括中科院计算所视觉信息处理与学习(Visual Information Processing and Learning, VIPL)研究组提出的局部Gabor幅相融合二值特征，并配合特征判别分析的方法[1, 2]。再如，在LFW数据集上，在允许利用有标签外部数据且非限定的测试条件下，性能最好的方法之一为微软亚洲研究院提出的关键特征点上的高维LBP特征，并配合稀疏回归判别特征提取的方法[3]。在上述测试条件下，该方法取得了95.17%的平均分类精度。

对LFW数据库而言，2014年是其性能得以戏剧性提升的一年。在2014年国际计算机视觉与模式识别会议(Conference on Computer Vision and Pattern Recognition 2014, CVPR 2014)上，

[1] Scale-invariant feature transform，尺度不变特征变换。
[2] Histogram of oriented gradient，方向梯度直方图。
[3] Local binary patterns，局部二值模式。

两个采用深度学习的团队——来自脸谱的团队[4]和香港中文大学的团队[5]，在允许利用有标签外部数据且非限定的测试条件下，分别报告了97.35%和97.45%的平均分类精度，比前述高维LBP特征方法的分类错误率降低了50%。上述两个团队均采用了卷积神经网络(Convolutional Neural Network, CNN)的变种架构。其中，脸谱的DeepFace方法强调前端的人脸3D对齐和虚拟正面化预处理，以削弱姿态变化的影响；而香港中文大学的DeepID方法则强调采用多个人脸区块分别训练卷积神经网络，并最终融合形成人脸特征表示。最近，该团队进一步开发了DeepID2+系统，在上述测试环境下取得了99.47%的正确分类精度，错误率比DeepID降低约80%。需要特别指出的是，这两个系统能够取得优异性能的另一个重要原因是均采用了大规模的标注人脸数据进行训练，而且其训练图像的分布与LFW测试图像（名人图像）有一定的相似性。例如，DeepFace采用了来自4030人的440万幅人脸图像（均来自社交网络）；而DeepID则使用了来自10177人的约20万人脸图像（均为网络名人图像）。

当然，在LFW上取得99.47%的正确分类精度并不代表人脸识别技术已经成熟。实际上，LFW数据集仅仅代表了人脸识别众多应用场景中的一种，即西方名人新闻照片识别。人脸识别还有很多其他应用场景，比如面向银行支付的人脸验证、面向智能视频监控的人脸识别等，尤其是后者，尚处于技术远远不能满足应用需求的状态。为了实现更为鲁棒和准确的识别，需要实现更为精确的面部特征定位，并处理好姿态、夸张表情和人脸老化等难题。

下面介绍近期我们在深度学习研究上的三个实践：利用大规模的人脸数据训练卷积神经网络并应用于基于视频的人脸和表情识别，由粗到精的深度非线性人脸形状提取方法，以及姿态鲁棒的人脸特征渐进深度学习方法。

用于人脸和表情识别的特征学习

人脸识别的最大预期应用是智能视频监控环境下的黑名单人物识别。此场景下的人脸识别距离实用还很遥远。尽管在完全真实视频监控场景下采集的公开人脸数据集尚不存在，但近两年出现了一些模拟类似场景的数据集。最近，美国NIST发布了PaSC视频人脸数据库。基于此数据集，贝弗里奇(R. Beveridge)等人组织了一个视频人物识别挑战评测。评测所用视频有两类：用固定摄像机拍摄和手持摄像机拍摄。视频拍摄时，活动中的人物并不是摄像机视野的焦点，因此，该数据库具有较大的挑战性。该评测在2014年生物特征识别国际联合会议(International Joint Conference on Biometrics 2014, IJCB'14)进行首次测试。针对手持摄像机视频的“视频-视频(video-to-video)”人脸验证场景，最好的算法为特征概率弹性匹配模型，其在错误接受率(False Accept Rates, FAR)为1%时的正确识别率（验证率）仅为26%，可见该测试之难。

最近，我们参加了第11届国际人脸和姿势自动识别会议(the 11th IEEE Conference on Automatic Face and Gesture Recognition, FG'15)组织的对PaSC的再次评测，所采用的技术方案有两个核心步骤：针对视频中每一帧人脸的卷积神经网络特征提取方法和集成视频片段中所有视频帧中人脸卷积神经网络特征的集合建模方法。所用卷积神经网络模型是在Caffe模型基础上改进得到的：我们将其从最初的5个卷积层增加为14个卷积层，并采用3个数据集对其训练：包括1520位名人的超过15万幅人脸图像，PaSC提供的训练集（共170人的约3.8万幅图像）和中科院计算所VIPL研究组创建的COX人脸库（共1000人的超过14万视频帧图像）。我们最终采用第二个全连接层的2048个隐节点输出作为单帧人脸图像的特征。在第二个步骤中，我们对每个视频片段中所有人脸帧的卷积神经网络特征进行集合建模，采用它们的均值、协方差和高斯分布作为视频表达，并分别学习这些集合表达的核测度，最终将每个视频

片段表达为1320维的视频特征进行比对。上述方法在FG'15的两个评测中均取得优秀的成绩[4]：针对固定和手持摄像机视频人脸识别，在错误接受率为1%时的验证率分别达到了58%和59%，显著高于第二名的48%和38%。与IJCB'14的最好结果相比，手持摄像视频情况下的识别率提高了125%。

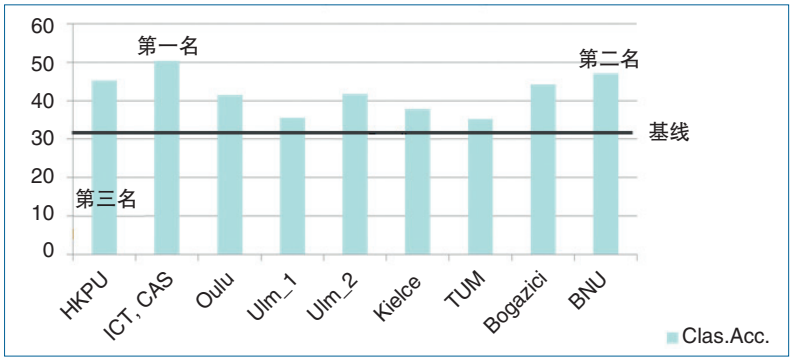


图1　ACM ICMI 2014举办的EmotiW'14竞赛的结果

此外，我们还尝试将卷积神经网络特征用于人脸表情识别。表情识别过去采用与人脸识别类似的局部特征（如HOG, SIFT, LBP, LGBP[5]等），并在此基础上用支持向量机等分类器进行分类。在第16届ACM国际多模式人机交互大会(the 16th ACM International Conference on Multimodal Interaction, ACM ICMI 2014)举办的EmotiW'14(the 2th Emotion Recognition In The Wild Challenge and Workshop)竞赛中，我们采用卷积神经网络模型学习人脸特征，将其与HOG和SIFT特征融合，并配合基于视频帧集合建模的方法[6]。我们的卷积神经网络模型是采用前述名人数据库（15万幅）训练的改进Caffe模型（7个卷积层），采用最后一个卷积层的12544个隐节点输出作为特征。需要特别注意的是，由于名人数据集缺少表情类别标签，该模型并不是针对表情识别精调的，而是针对人脸识别精调的。即便如此，该卷积神经网络模型也取得了与SIFT特征基本相同的性能（在校验集上表情识别率分别为43.40%和43.94%），均优于HOG特征(38.01%)。将这三种特征通过三种集合模型（子空间、协方差和高斯分布）和三种分类器（支持向量机、逻辑回归分类器[6]和偏最小二乘分类器[7]）进行融合，在校验集合上识别率提高到45.28%，与音频特征融合后提高到48.52%。最终在主办方的盲测试集合上取得了50.37%的识别率，是9个参赛算法中最高的（见图1）。

多阶段深度非线性人脸形状提取

面部特征点定位（又称人脸形状提取或人脸对齐）在人脸识别、表情识别、人脸动画合成等诸多任务中具有非常重要的作用。由于姿态、表情、光照和遮挡等因素的影响，真实场景下的

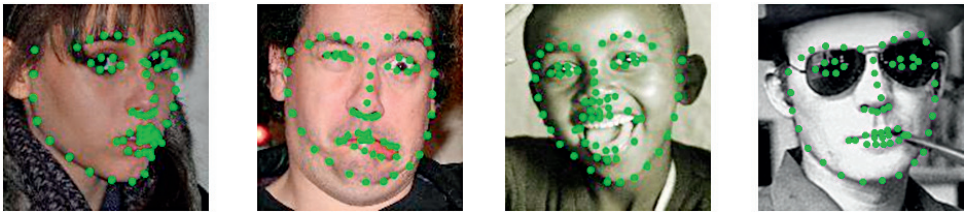


图2　姿态、表情、光照和遮挡条件下的面部特征点定位

[4] 目前评测的最终结果尚未正式发布，但组织者发布的信息表明我们的结果显著优于其他参赛者。

[5] Local gabor binary pattern，局部Gabor二值模式。

[6] Logistic Regression Classifier, LR。

[7] Partial Least-Square, PLS。

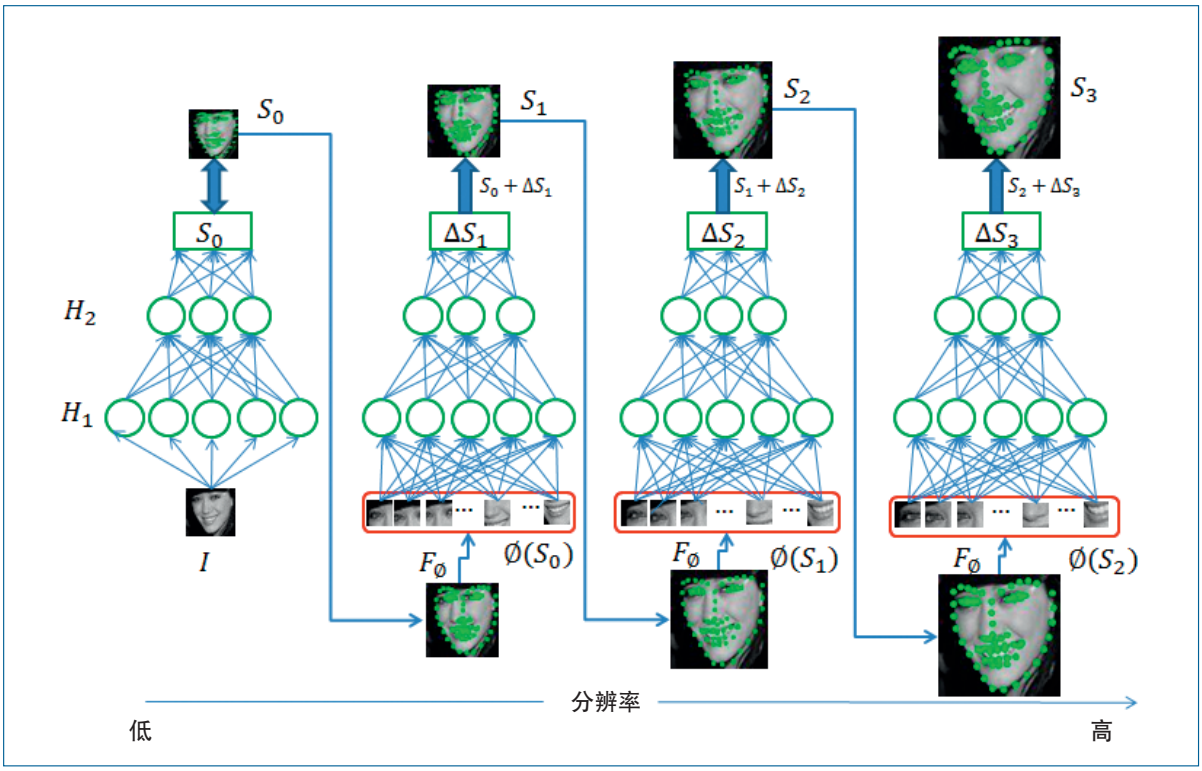


图3　多阶段深度非线性人脸形状提取方法

人脸对齐任务是一个非常困难的问题（见图 2）。

主动形状模型 ASM 和主动表观模型 AAM 是经典的人脸对齐方法，它们使用线性的主成分分析技术对人脸形状和纹理变化建模，并通过优化模型参数使之适配测试人脸图像。由于线性模型难以刻画复杂的人脸形状和纹理变化，在大姿态、夸张表情、剧烈光照变化和部分遮挡下的效果欠佳。解决该问题的最新进展是通过级联多个线性回归模型直接从人脸纹理特征（改进的 SIFT）预测人脸形状。为进一步提高算法的非线性回归能力以获得对姿态等变化的鲁棒性，我们提出了一种由粗到细的深度非线

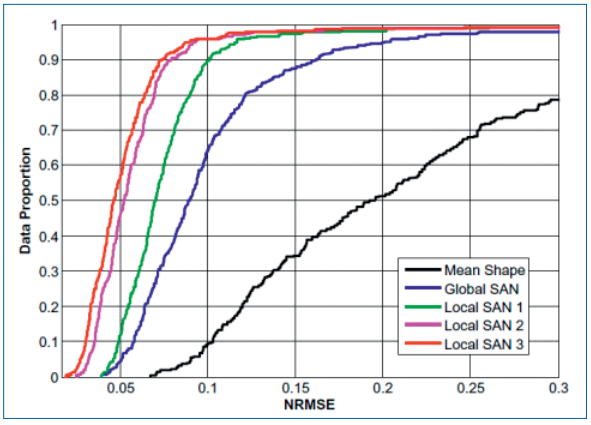


图4　每级自编码器网络性能增益

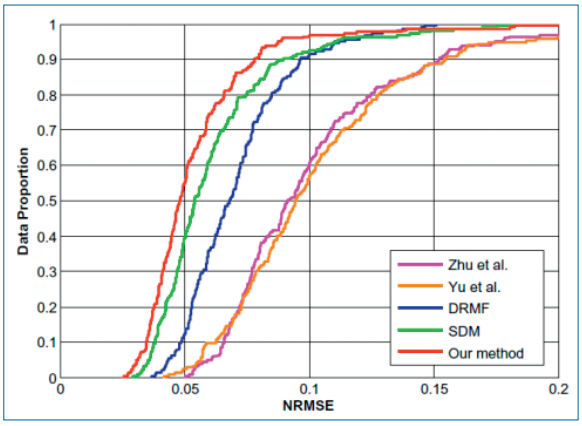


图5　LFPW数据集性能对比

性人脸形状提取方法——CFAN[8]。CFAN 级联多个由栈式自编码网络实现的非线性回归模型（见图 3），每一级刻画从人脸表观到人脸形状的部分非线性映射。

CFAN 第一级栈式自编码网络 (SAN) 以较低分辨率的人脸图像作为输入，快速估计出粗略的人脸形状 S_0。然后在分辨率更高的人脸图像上提取 S_0 各特征点的 SIFT 特征，作为下一级自编码网络的输入来优化人脸形状得到 S_1。以此类推，我们级联多个基于局部特征的自编码网络，在分辨率不断提高的人脸图像上逐步优化人脸形状提取结果。该方法中，每个自编码网络刻画了部分“非线性”，多个级联有效逼近了全局“非线性”[7]。

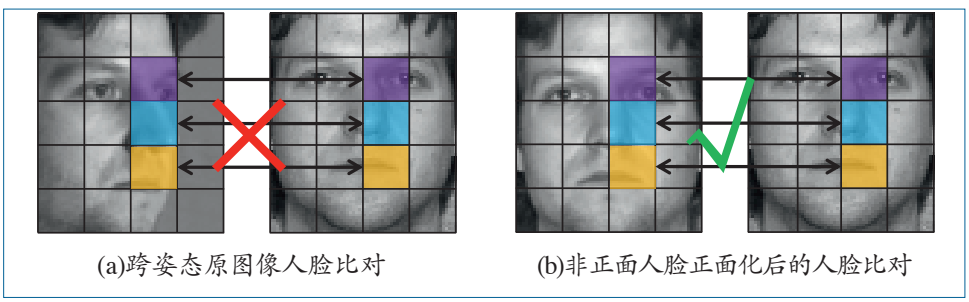
我们使用 3 个公开的数据集——LFPW, HELEN 和 AFW 来验证 CFAN 方法的有效性。我们合并 LFPW 的训练集合、HELEN 和 AFW 作为训练集进行模型训练，并在 LFPW 的测试集上进行测试。图 4 展示了在 LFPW 测试集上各级自编码网络的面部特征点定位正确率。不难看出，后续自编码网络的特征定位性能逐步增强。与 SDM 等最先进方法的对比结果如图 5 所示，可见我们的方法取得了最佳的定位性能。

姿态鲁棒的人脸特征渐进深度学习

人脸识别技术在可控条件及半可控条件下已经基本趋于成熟，然而在非可控条件下，由于受姿态、光照、表情、年龄等因素的影响，人脸识别依然很不成熟。其中，姿态变化会导致极大的面部表观变化，是对人脸识别影响最大的因素之一。

姿态变化引起的面部表观变化使不同人脸图像相同像素位置的语义不同。如图 6(a) 中蓝色块所示，正面像代表的是一半鼻子，而侧面像则包含了整个鼻子，直接进行图像比较与识别通常性能很差。为了解决跨姿态的人脸识别问题，已有的工作主要从两个方面进行，即姿态鲁棒的特征提取方法和生成虚拟正面人脸图像的方法。前者一般通过对不同姿态图像之间的映射关系或公共特征表示建模来实现。后者则通常利用 3D 模型或者 2D 学习的方式生成非正面人脸图像的虚拟正面人脸图像，如图 6(b) 所示，从而使不同姿态的图像可以在相同姿态下进行比较与识别。

姿态变化导致的人脸表观变化是一种复杂的非线性变化，利用 3D 模型生成虚拟图像的方式固然可以较好地解决不同姿态间



(a)跨姿态原图像人脸比对　　(b)非正面人脸正面化后的人脸比对

图6　跨姿态人脸识别难点及虚拟正面人脸方法

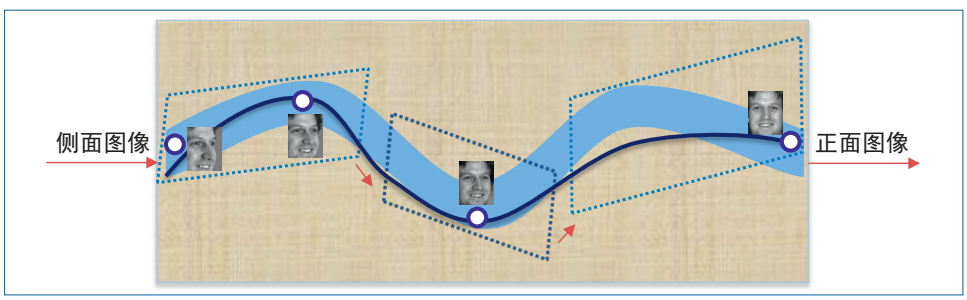


图7　侧面图像到正面图像的栈式渐进非线性建模示意图

[8] Coarse-to-Fine Auto-encoder Networks，由粗到精自动编码网络。

表1 在MultiPIE数据库上的性能对比

对比方法		测试角度							姿态估计
		-45°	-30°	-15°	+15°	+30°	+45°	平均	
3D方法	3D Pose Norm(Asthana et al. ICCV 2011)	0.741	0.910	0.957	0.957	0.895	0.748	0.868	自动估计
	MDF(Li et al. ECCV 2012)	0.787	0.940	0.990	0.987	0.922	0.818	0.907	
2D方法	GMA (Sharma et al. CVPR 2012)	0.750	0.745	0.827	0.926	0.875	0.652	0.796	人工设定
	DAE（直接应用自编码器）	0.699	0.812	0.910	0.919	0.865	0.743	0.825	不需要
	SAPE（我们的方法）	0.849	0.926	0.963	0.957	0.943	0.844	0.914	

的非线性变化问题，但从 2D 图像恢复准确的 3D 模型非常困难。考虑到深度神经网络有很强的非线性建模能力，可以采用深度学习来建模。然而，深度学习需要大规模的有监督、多姿态人脸图像进行训练，而这类数据在实际中很难收集。为此，我们提出一种栈式渐进自编码 (SPAE) 神经网络模型，以实现较小规模数据下对姿态变化的非线性建模。侧面图像到正面图像变化虽然非常复杂，但却是缓慢平滑的，如图 7 所示。根据这一特点，我们将侧面图像到正面图像的建模划分为若干子任务，每个子任务仅负责将变化较大的姿态变换到变化较小的姿态而非直接变换到正面姿态，由此控制了每个子问题的难度，使用一个浅层的神经网络即可有效建模，进而将多个浅层的神经网络叠联到一起即可得到一个深层的神经网络，实现侧面图像到正面图像的平滑变换。这种渐进学习的思想将深度神经网络划分为若干浅层的网络，使其模型能力与有限的数据相匹配，避免数据规模小带来的过学习问题。

每个子任务，亦有多种网络结构可供选择。我们采用自编码器网络并对其进行改进以适应栈式渐进的需求。如图 8 所示，我们将每个浅层自编码器的目标设计为仅进行较小范围的姿态转化，即将变化较大姿态的图像转换到相邻的变化较小姿态，而姿态变化已经较小的图像则保持不变。由此通过级联多个浅层的渐进自编码器形成的深层网络结构即可实现平滑的姿态变换。随着网络层数的增加，姿态变化越来越小。在最高层，所有图像均被转换为正面姿态。此网络中高隐层包含较小的姿态变化，其隐层输出即可用作姿态鲁棒的特征 [8]。

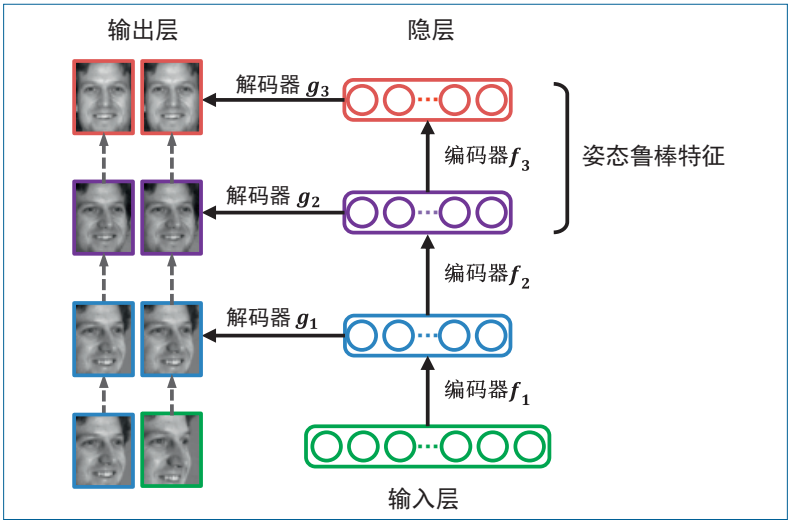


图8 栈式渐进自编码网络示意图

我们在 MultiPIE 数据库上对上述方法进行了评测，结果如表 1 所示。其中，0° 姿态为目标

集，其余6个姿态用作测试集。从识别结果可以看出，我们提出的SPAE方法的性能明显优于其他2D方法，甚至略优于基于3D的方法。需要指出的是，该方法的另外一个重要优点是不需要已知输入图像的姿态，也不需要进行显式的姿态估计。

总结与讨论

本文概述了我们应用卷积神经网络进行人脸处理和识别的近期进展，印证了卷积神经网络等深度学习方法在大规模人脸数据训练条件下所能取得的优异性能。我们提出了两个采用栈式自编码深度网络的方法，即多阶段深度非线性人脸形状提取方法和姿态鲁棒的人脸特征渐进学习方法。这两个方法中，我们并非简单地应用栈式自编码深度网络，而是针对具体问题进行了重新设计。在此，我们将其中的经验总结如下：

1. 深度学习模型的推广性能依赖于大规模训练数据的支持。在训练数据规模足够大的时候，直接应用现有深度学习模型往往可以取得优越的性能。但多大规模的数据是“足够”的，在理论和实践上均有待深入研究。

2. 在训练数据较少的情况下，直接应用深度学习模型难以取得预期的性能。例如，在本文所述的人脸形状提取和姿态鲁棒特征提取问题上，从图4中Global SAN的性能以及表1中DAE的性能都不难看出这样的规律。

3. 在训练数据较少的情况下，我们通过引入先验知识适应性的调整深度模型，利用“由粗到精”的思想将原问题分解为多个相对简单的子问题，分别对应深度网络每一层的优化目标，从而降低了每一层训练所需的数据规模，取得了令人满意的效果。■

致谢：

除本文作者外，中国科学院计算技术研究所VIPL研究组的博士生黄智武、刘梦怡、李岩以及副研究员王瑞平均在FG'15和EmotiW'14竞赛中作出了重要贡献，特此感谢他们对本文相关工作的支持。

山世光

CCF会员、本刊特邀专栏作家。中科院计算所研究员。主要研究方向为计算机视觉、模式识别和机器学习。sgshan@ict.ac.cn

阚美娜

CCF会员。中科院计算所助理研究员。主要研究方向为计算机视觉和模式识别。kanmeina@ict.ac.cn

李绍欣

中科院计算所博士生。主要研究方向为计算机视觉和模式识别。shaoxin.li@vipl.ict.ac.cn

其他作者：张　杰　陈熙霖

参考文献

[1] S. Xie, S. Shan, X. Chen, J. Chen. Fusing Local Patterns of Gabor Magnitude and Phase for Face Recognition. *IEEE Trans. on Image Processing*. 2010; 19(5): 1349~1361.

[2] Y. Li, S. Shan, H. Zhang, S. Lao, X. Chen. Fusing Magnitude and Phase Features for Robust Face Recognition. ACCV 2012.

[3] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification. CVPR, 2013.

[4] Y. Taigman, M. Yang, M. Ranzato, L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. CVPR 2014.

[5] Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation from Predicting 10000 Classes. CVPR 2014.

[6] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, X. Chen. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. ACM ICMI 2014, Nov. 2014.

[7] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-Fine Auto-encoder Networks (CFAN) for Real-time Face Alignment. ECCV 2014.

[8] M. Kan, S. Shan, H. Chang, X. Chen. Stacked Progressive Auto-Encoder (SPAE) for Face Recognition Across Poses. CVPR 2014.