DeepFace: Closing the Gap to Human-Level Performance in Face Verification

DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman Ming Yang Marc'Aurelio Ranzato Lior Wolf

Facebook AI Research
Menlo Park, CA, USA
{yaniv, mingyang, ranzato}@fb.com

Tel Aviv University
Tel Aviv, Israel
wolf@cs.tau.ac.il

摘要

传统的人脸识别流程是：**人脸检测——人脸对齐——人脸表达——人脸分类。**
为了进行分段的仿射变换，我们使用了3D的人脸建模来重现对齐和表达这两步，最终从一个9层的深度神经网络中得到了人脸的表达。
这个网络并非标准的卷积网络层，而是使用了几个未共享权重的局部连接层，网络参数超过了120，000，000个。
我们在迄今为止最大的人脸数据库上训练——4000多个不同的人，总计440万张带标记的人脸库。
这种在大型数据库中基于模型进行准确的对齐并用神经网络训练学习到的人脸表达，可以很好地推广到非受限环境下的人脸表达。
我们的方法在LFW上达到了97.35%的人脸验证精度，逼近了人类的水平。

1 介绍

非受限条件下的人脸识别是认知领域的前沿算法。人脸识别技术对社会文化的影响是深远的，
然而当前在这一领域机器和人类的认知能力差距是我们不得不面对的现实。
我们提出了一个系统——DeepFace。它已经缩小了现存的在非受限条件下进行人脸识别的差距，已经达到了人类在此领域的认知水平。
本系统是在一个极大人脸库上训练的，并且这个训练库与测试评估所用的库是截然不同的，并且可以轻松超越现存的其他系统。
更进一步，与过去的成百上千种基于外观特性的系统相比，本系统产生的是一个很简洁的人脸特征表达。

本系统与大多数传统做法相比最大的区别在于它使用深度学习，代替了人工设计的特征。深度学习特别适合解决大样本训练集，近期它在不同领域也都获得了成功，如计算机视觉，演讲和语音建模等领域。
特别的，针对人脸时，这种训练学习到的神经网络在提取人脸特征上的成功高度依赖于快速的3D对齐的步骤。网络结构是基于这种假定：一旦完成了对齐，人脸局部的每个区块在像素级别上就固定了，因此才可能从RGB值进行学习，而不需像其他系统【19，21】得应用好几个卷积层。

我们做出以下工作：
1，一种有效的深度神经网络结构的构建，实现了利用大型带标签数据库进行训练学习获得人脸表达的方法，并使其能很好的泛化到其他数据库上；
2，基于3D人脸模型设计的一个有效的人脸对齐系统；
3，极大地提升了在LFW库上机器认知性能，近乎达到了人类的性能，在YouTube人脸库(YTF)上，将错误率降低了超过50%。

1.1 相关工作

大数据和深度学习
近年来，网络用户使用搜索引擎爬取了大量图片，并且上传到社交媒体，这里面包含了大量非受限条件下的数据，如各类物体，人脸和风景。

越来越多的数据和不断增长的计算资源使得使用更强劲的统计模型成为可能。这些模型已经大幅度提升了计算机视觉系统在几个重要方面的鲁棒性：如非刚性形变，混杂，遮挡和光照变化。而这些问题往往是计算机视觉应用的难点。

传统的机器学习方法，如支持向量机，主成分分析和线性鉴别分析。在利用大数据方面有明显的容量上限，而深度神经网络则表现出很好的大规模训练特性。

近期，神经网络方向已经吸引了许多研究学者。尤其当大型神经网络展现出优良的结果时：1，可以使用神经网络训练大量的数据；2，大规模的计算资源（大量的CPU或者GPU）更容易获得。【19】也证明了在一个大的数据库上使用标准的反向传播算法训练的大型卷积网络可以实现很高的识别精度。

人脸识别的最高水平
近几十年来，在受限条件下的人脸识别准确度不断提高，许多商业公司也在受限的条件下，部署开发了一些生物识别系统，然而这些系统在几个方面的鲁棒性很差，如光照变化，表情变化，遮挡和年龄变化。这导致在非受限条件下，这些系统的性能急剧恶化。

现阶段很多人脸验证方法都使用人工设计的特征，并且这些特征通常被结合起来使用以提升性能。当时比较领先的系统甚至使用了成千上万个图像描述子。相反的，我们的方法仅仅使用RGB的像素值，就能产生非常简洁且稀疏的描述子。

深度神经网络在过去也被用在人脸检测，人脸对齐和人脸验证等方面。在非受限条件下，【16】使用LBP特征作为输入，他们证明结合传

统方法可以能带来性能提升。在我们的系统中，我们使用未加工的图像信息作为底层的输入，避免结合工程设计的特征描述子，以此来强调神经网络自身的贡献。

度量学习方法在人脸验证领域应用较广，且通常结合具体的对象。文献【5】在一个大型的标记数据库上，采用适应联合贝叶斯模型【6】的转移学习方法，学习到数据库A(2995个人，99773个人脸)到LFW的转移模型。为了证明本系统提取的特征的高效性，我们做人脸验证时仅使用简单的距离度量。

人脸对齐

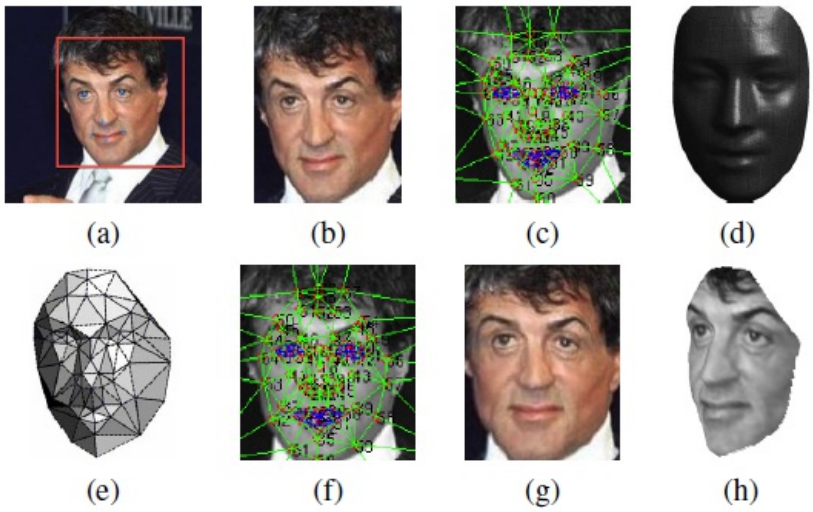

Figure 1. **Alignment pipeline.** (a) The detected face, with 6 initial fiducial points. (b) The induced 2D-aligned crop. (c) 67 fiducial points on the 2D-aligned crop with their corresponding Delaunay triangulation, we added triangles on the contour to avoid discontinuities. (d) The reference 3D shape transformed to the 2D-aligned crop image-plane. (e) Triangle visibility w.r.t. to the fitted 3D-2D camera; darker triangles are less visible. (f) The 67 fiducial points induced by the 3D model that are used to direct the piece-wise affine warpping. (g) The final frontalized crop. (h) A new view generated by the 3D model (not used in this paper).

3.表达

这些年来，在计算机视觉表达方面，有很多研究贡献，这种应用与人脸识别的描述子，多数是在人脸图像的所有位置进行相同的运算（LBF等）。最近，由于大量的数据更容易获得，基于学习的方法变得比较流行,因为这种方法针对特定的任务，发掘和最优化相应的特征。本文中，我们通过一个大型的深度网络学习人脸图像的通用表达。

DNN的结构和训练

我们通过一个多类人脸识别任务来训练深度神经网络（DNN）,也即认清一张人脸所属身份的问题。整体的网络结构如图2所示：

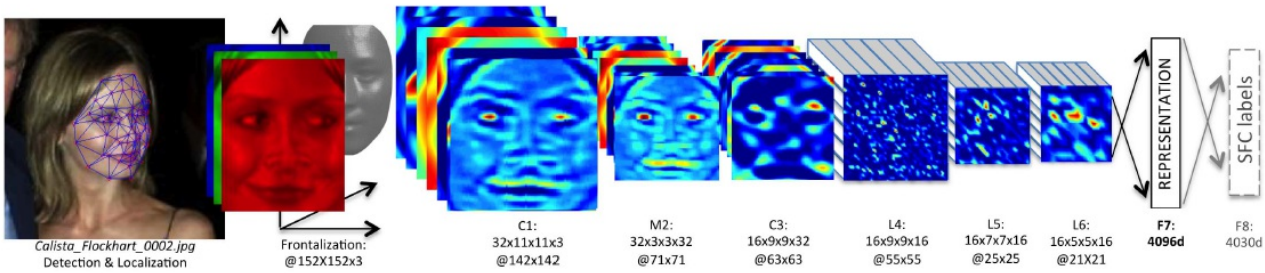

Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

1，预处理阶段：输入3通道的人脸，并进行3D校正，再归一化到152*152像素大小——152*152*3。
2，通过卷积层C1：C1包含32个11*11*3的滤波器（即卷积核），得到32张特征图——32*142*142*3。
3，通过max-polling层M2：M2的滑窗大小为3*3，滑动步长为2，3个通道上分别独立polling。
4，通过另一个卷积层C3：C3包含16个9*9*16的3维卷积核。
上述3层网络是为了提取到低水平的特征，如简单的边缘特征和纹理特征。Max-polling层使得卷积网络对局部的变换更加鲁棒。如果输入是校正后的人脸，就能使网络对小的标记误差更加鲁棒。然而这样的polling层会使网络在面部的细节结构和微小纹理的精准位置上丢失一些信息。因此，我们只在第一个卷积层后面接了Max-polling层。这些前面的层我们称之为前端自适应的预处理层级。然而对于许多计算来讲，这是很必要的，这些层的参数其实很少。它们仅仅是把输入图像扩充成一个简单的局部特征集。

后面的层：

5，L4,L5,L6都是局部连接层，【13，16】，就像卷积层使用滤波器一样，在特征图像的每一个位置都训练学习一组不同的滤波器。由于校正后不同区域的有不同的统计特性，卷积网络在空间上的稳定性的假设不能成立。比如说，相比于鼻子和嘴巴之间的区域，眼睛和眉毛之间的区域展现出非常不同的表观并且有很高的区分度。换句话说，通过利用我们输入的校正后的图像，我们定制了DNN的结构。

使用局部连接层并没有影响特征提取时的运算负担，但是影响了训练的参数数量。仅仅是由于我们有如此大的标记人脸库，我们可以承受三个大型的局部连接层。局部连接层的输出单元受到一个大型的输入图块的影响，可以据此调整局部连接层的使用（参数）（不共享权重）

比如说，L6层的输出受到一个74*74*3的输入图块的影响，在校正后的人脸中，这种大的图块之间很难有任何统计上的参数共享。

6，最后，网络顶端的两层（F7，F8）是全连接的：每一个输出单元都连接到所有的输入。这两层可以捕捉到人脸图像中距离较远的区域的特征之间的关联性。比如，眼睛的位置和形状，与嘴巴的位置和形状之间的关联性（这部分也含有信息）可以由这两层得到。第一个全连接层F7的输出就是我们原始的人脸特征表达向量。

在特征表达方面，这个特征向量与传统的基于LBP的特征描述有很大区别。传统方法通常使用局部的特征描述(计算直方图)并用作分类器的输入。

最后一个全连接层F8的输出进入了一个K-way的softmax（K是类别个数），即可产生类别标号的概率分布。用Ok表示一个输入图像经过网络后的第k个输出，即可用下式表达输出类标号k的概率：

$$p_k = \exp(o_k) / \sum_h \exp(o_h)$$

训练的目标是最大化正确输出类别（face 的id）的概率。我们通过最小化每个训练样本的叉熵损失实现这一点。用k表示给定输入的正确类别的标号，则叉熵损失是：$L = -\log p_k$

通过计算叉熵损失L对参数的梯度以及使用随机梯度递减的方法来最小化叉熵损失。

梯度是通过误差的标准反向传播来计算的【25，21】。非常有趣的是，本网络产生的特征非常稀疏。超过75%的顶层特征元素是0。这主要是由于我们使用了ReLU激活函数导致的。这种软阈值非线性函数在所有的卷积层，局部连接层和全连接层（除了最后一层F8）都使用了，从而导致整体级联之后产生高度非线性和稀疏的特征。稀疏性也与使用使用dropout【19】正则化有关，即在训练中将随机的特征元素设置为0。我们只在F7全连接层使用了dropout.由于训练集合很大，在训练过程中我们没有发现重大的过拟合。

给出图像I,则其特征表达G(I)通过前馈网络计算出来，每一个L层的前馈网络，可以看作是一系列函数g_ϕ^l构成。然后表达成：$G(I) = g_\phi^{F_7}(g_\phi^{L_6}(...g_\phi^{C_1}(T(I,\theta_T))...))$，其中网络参数$\phi = \{C_1,...,F_7\}$，$\theta_T = \{x_{2d}, \vec{P}, \vec{r}\}$表示图像的姿态等信息。

归一化

在最后一级，我们把特征的元素归一化成0到1，以此降低特征对光照变化的敏感度。特征向量中的每一个元素都被训练集中对应的最大值除。然后进行L2归一化。由于我们采用了ReLU激活函数，我们的系统对图像的尺度不变性减弱。

4，验证

4.1加权的χ^2距离

本系统中，归一化后的DeepFace特征向量与传统的基于直方图的特征（如LBP）有一下相同之处：

1，所有值均非负；

2，非常稀疏；

3，特征元素的值都在区间[0，1]之间。

因此，我们使用加权的χ^2相似度：$\chi^2(f_1,f_2) = \sum_i w_i(f_1[i] - f_2[i])^2/(f_1[i] + f_2[i])$

其中，f1,f2是DeepFace特征。权重参数用线性SVM学习得到（将$(f_1[i] - f_2[i])^2/(f_1[i] + f_2[i])$作为整体训练SVM）。

4.2 Siamese network(一个监督的度量学习模型)

我们也尝试了一个端到端的度量学习方法，即Siamese network【8】：一旦学习（训练）完成，人脸识别网络（截止到F7）在输入的两张图片上重复使用，将得到的2个特征向量直接用来预测判断这两个输入图片是否属于同一个人。这分为以下步骤：a,计算两个特征之间的绝对差别；b,一个全连接层，映射到一个单个的逻辑单元（输出相同/不同）。

$d(f_1,f_2) = \sum_i \alpha_i |f_1[i] - f_2[i]|$.

α_i是Siamese网络参数。该网络也是采用最小化叉熵损失和误差反向传播算法训练得到。

5 实验

5.1 数据集

训练数据集：**SFC**，社交人脸分类数据库。来自facebook。
测试数据集：**LFW**,现在基准的非受限人脸验证数据库；
YTF,来自YouTube,与LFW属性相似，主要是视频截图。

SFC:
4400000张带标记的人脸，含有4030个人，每个人拥有大约800-1200张人脸。每个人最近期的图片的5%留下来用于测试（通过照片的拍摄时间切分），这样做是为了通过年龄模仿连续的验证；拥有大量图片的人物可以满足学习不变性的需求。我们通过几种自动的方法进行验证。通过检查几个训练集中人物的名字，我们避免了几个数据库拥有同一个人的情况。

LFW:
13323张网络图片，包含5749个名人。分为6000个人脸对（共10组），

YTF:
1595个人的3425段视频（LFW中的人）。将其分为5000个视频对（10组）,用来评估视频级别的人脸验证。

SFC中的人物身份是人工标注的，**典型的标注误差是3%**。社交图片在图片质量，光照和表情更加参差不齐，而LFW和YTF通常是名人并且是由专业的摄像设备拍摄。

5.2 在SFC上训练

我们首先在SFC上使用多分类问题训练深度神经网络（基于GPU,通过SGD实现前馈网络上的bp算法）。
最小的batch大小(一批图像)是128个，所有训练的网络层的学习率都是0.01，当验证错误率停止减少时，手动将LR按数量级降低，只到最终达到0.0001。
每一层网络的初始权重按照0均值的高斯分布来初始化（ $\sigma = 0.01$ ，偏置为0.5）。
耗时3天

在SFC的5%的数据集上，我们评估了不同DNN设计下的分类误差。这验证了使用大型人脸数据库和深度网络结构的必要性。
1，首先对SFC数据库按照每个人物的图片子集进行划分。分别选取了1.5K，3K，4K个人（分别总计1.5M,3.3M,4.4M个人脸）。按照上面的网络结构，我们训练了3个神经网络，DF-1.5K,DF-3.3K,DF-4.4K
2，改变总样本量，依次为SFC总量的10%，20%，50%（即每个人的图片量按比例取样本），训练3个DNN,SFC-10%,SFC-20%,SFC-50%。
3，改变网络结构，分别去除C3，L4+L5，C3+L4+L5。在4.4M的SFC库上训练得到3个DNN,DF-sub1,DF-sub2,DF-sub3。DF-sub3具有最浅的结构。

Network	Error	Network	Error	Network	Error
DF-1.5K	7.00%	*DF-10%*	20.7%	*DF-sub1*	11.2%
DF-3.3K	7.22%	*DF-20%*	15.1%	*DF-sub2*	12.6%
DF-4.4K	8.74%	*DF-50%*	10.9%	*DF-sub3*	13.5%

Table 1. Comparison of the classification errors on the *SFC w.r.t.* training dataset size and network depth. See Sec. 5.2 for details.

结论：
1，当训练人数规模提高的时候，分类误差只是稍有变化，这证明了网络可以负载大规模人物的训练集。
2，当参与训练的图片总量减少时，分类误差升高到20.7%，这是因为训练集聚减后，出现了过拟合。证明训练集总数越大，网络性能越好。
3，当网络结构精简时，层数越少的网络的分类误差最终会更大。这证明了在大型人脸数据集上训练时，网络深度很重要。

5.3 LFW上的测试结果

人类在LFW库上的人脸验证准确度为97.5%。DeepFace使用3D校正的人脸完成基于前馈的大型网络。针对本系统的各部分功能：
1，倘若只用2D的校正，准确率为94.3%，完全不校正，准确率为87.9%；
2，不用神经网络，用3D校正结合朴素LBP/SVM,达到91.4%。

所有的LFW中的图片的处理过程与用SFC图片训练时一样，为了独立地评估本系统人脸表达的区分度能力，我们遵循无监督的做法，直接比较归一化之后的特征对的内积。非常显著的，这达到了95.92%的准确率，基本与当前最领先的有监督转移学习方法【5】相当。
接着，我们在 χ^2 距离向量的基础上学习了一个SVM，在5400对带标记的LFW人脸对上测试，达到了97%的准确率，极大地降低了错误率。

Method	Accuracy ± SE	Protocol
Joint Bayesian [6]	0.9242 ±0.0108	restricted
Tom-vs-Pete [4]	0.9330 ±0.0128	restricted
High-dim LBP [7]	0.9517 ±0.0113	restricted
TL Joint Bayesian [5]	0.9633 ±0.0108	restricted
DeepFace-single	**0.9592** ±0.0029	unsupervised
DeepFace-single	**0.9700** ±0.0028	restricted
DeepFace-ensemble	**0.9715** ±0.0027	restricted
DeepFace-ensemble	**0.9735** ±0.0025	unrestricted
Human, cropped	0.9753	

Table 3. Comparison with the state-of-the-art on the *LFW* dataset.

DNN的整体效果

接着，通过调整不同类型的输入。我们得到了多个DNN网络。

1，输入3D校正后的RGB图像——DeepFace-single；

2，输入灰度图加上图像梯度和方向等信息——DeepFace-gradient；

3，输入2D校正后的RGB图像——DeepFace-align2d.

我们使用基于CPD核的非线性SVM来将这些距离度量结合起来

（ $K_{\text{Combined}} := K_{\text{single}} + K_{\text{gradient}} + K_{\text{align2d}} \quad K(x,y) := -||x-y||_2$ ）

达到97.15%的准确率。

又使用了100K个新的人物，每人30张图片作为样本训练了Siamese网络。并将Siamese网络与上述网络结合起来，$K_{\text{Combined}} \mathrel{+}= K_{\text{Siamese}}$，这样将准确率提高到97.25%

又额外增加了4个DeepFace-single网络，$K_{\text{Combined}} \mathrel{+}= \sum K_{\text{DeepFace-Single}}$ 将准确率提高到97.35%。

Network	Error (*SFC*)	Accuracy ± SE (*LFW*)
DeepFace-align2D	9.5%	0.9430 ±0.0043
DeepFace-gradient	8.9%	0.9582 ±0.0037
DeepFace-Siamese	NA	0.9617 ±0.0038

Table 2. The performance of various individual *DeepFace* networks and the Siamese network.

5.4 YTF数据库上的结果

进一步在近期的视频级人脸验证数据库上验证我们的DeepFace。YTF视频帧的图像质量比webt图片的质量更差（运动毛刺和远距离拍摄等因素）。每个视频对挑出，50个视频帧对，并且根据视频源的名字对齐进行标注（一个人/不是一个人），然后训练网络。给出一个测试视频对后，从每段视频中随机选出100个视频帧对，将输出的结果取均值作为判断的依据。

在YTF上，我们得到了91.4%的准确率，由于YTF库中有100个标注错误的视频对，经过改正后，我们的准确率达到了92.5%。这也证明我们的DeepFace方法，在其他领域也具有很好的泛化性能（视频人脸验证）。

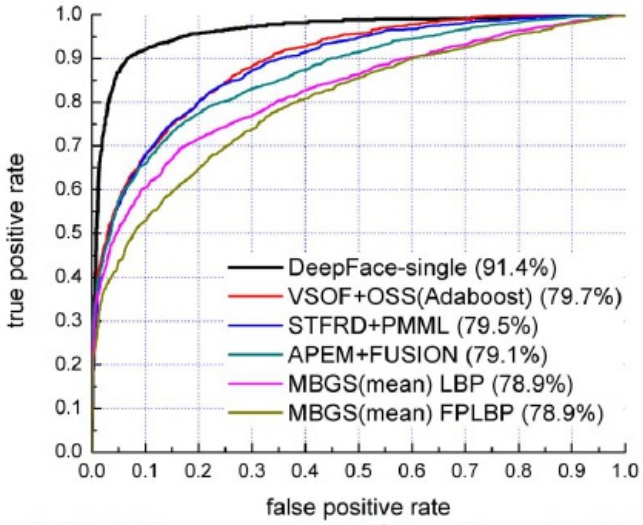

Figure 4. The ROC curves on the *YTF* dataset. Best viewed in color.

5.5计算效率

单核Intel 2.2GHz CPU，从原始输入像素中提取特征的时间是180ms，3D校正时间是50ms，每张图片的处理总时间是330ms=图像解码+人脸检测+校正+应用前馈网络+输出最终分类。