从头开始学习面部表征

Dong Yi, Zhen Lei, Shengcai Liao and Stan Z. Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition

Institute of Automation, Chinese Academy of Sciences (CASIA) dong.yi, zlei, scliao, szli@nlpr.ia.ac.cn

Abstract

通过大数据和深度卷积神经网络（CNN）推动，人脸识别的性能正在变得与人类相当。 使用私人大规模训练数据集，若干组在LFW上实现了非常高的性能，即97％至99％。 虽然有许多CNN的开源实现，但没有一个大规模的人脸数据集是公开可用的。 人脸识别领域的现状是数据比算法更重要。 为了解决这个问题，本文提出了一种从Internet上收集人脸图像的半自动方法，并建立了一个包含大约10,000个主题和500,000个图像的大规模数据集，称为CASIAWebFace。 基于数据库，我们使用11层CNN来学习判别表示并获得LFW和YTF的精确状态。 CASIAWebFace的出版将吸引更多的研究团体进入这一领域，并加速野外人脸识别的发展。

# Introduction

在过去的一年中，人脸识别算法的性能大幅提升。 例如，LFW [7]（目前最难的面部数据集）的准确度从95％提高到99％[24]，这与人类表现相当。 LFW的最佳方法可分为两类：宽模型和深模型。 良好模型的本质是它应该具有足够的能力来表示复杂人脸图像的变化。 高维LBP [5]是一种典型的宽模型，它通过将图像转换为非常高维的空间来平整面部歧管。 卷积神经网络（CNN）[12]是用于人脸识别和图像分析的最先进的深度模型。

arXiv:1411.7923v1 [cs.CV] 28 Nov 2014

虽然模型可以沿“宽度”或“深度”方向增加其复杂性，但是当具有相同数量的参数时，深模型比宽模型更有效[2]。而且，普通电脑不能轻松1

ily处理宽模型提取的高维特征。相反，每层深层模型的特征维数都要低得多，这使得深层模型的内存消耗成本低廉。 CNN从数据中提取知识的力量已在may领域得到验证[10]。最近，深度CNN正在成为人脸识别的主流，并在LFW上占据主导地位。然而，受到培训数据规模的限制，CNN的上限尚未测量。

CNN的许多优秀的开源实现[9,8]可用于从大数据中学习人脸表示，但直到现在还没有研究团队将其私有大型数据集公开。原因可能是大型数据集很难收集，需要消耗大量资金和人力。但对于学术研究而言，开发私有数据算法在两个方面是有害的：首先，大多数研究人员无法为缺乏数据的大规模人脸识别方法做出贡献。其次，由于训练集的不同，许多经典方法与CNN无法比较。

在LFW上，除了LFW之外，最好的方法都使用外部数据，即LFW的“Unrestricted，Labeled Outside Data”类别。事实上，这类方法很难简单地称为方法，而是至少包括外部数据和算法的解决方案。为了扩展LFW的规模并标准化“无限制，标记的外部数据”的评估协议，本文构建了一个包括大约10,000个主题和500,000个图像的大规模数据集，称为CASIA-WebFace。据我们所知，该数据集的大小在文献中排名第二，仅小于Facebook的私有数据集（SCF）[26]。我们鼓励这些数据消费方法对此数据集进行培训并报告LFW的性能。

从互联网爬行的脸部图像很容易，但注释他们的身份很难。由于IMDb网站的良好结构，爬行和注释可以半自动方式完成。首先，从网站上抓取一些感兴趣的名人的名字，然后下载他们页面中的照片。因为大多数照片通常包含多个面部，所以难以出现。因此，我们提出了一种简单快速的聚类方法来注释照片中人脸的身份。为确保数据集中的主题不与LFW重叠，我们使用名称的编辑距离来检查重复。最后，我们通过手动扫描整个数据集并更正错误注释。

为了说明CASIA-WebFace的质量，我们在其上训练了一个深度CNN。我们的网络集成了最近作品中最受欢迎的组件，如ReLU neuron [20]，dropout [6]，低维表示，识别+验证成本函数[24]，小滤波器和非常深的架构[23]。我们的网络根据标准协议和新提出的协议“BLUFER”[17]在LFW上进行评估。所有实验结果均表明其优越的性能。该网络还在另一个具有挑战性的数据集YouTube Faces（YTF）[28]上进行了测试。使用从CASIA-WebFace获得的表示，我们获得了与Facebook的DeepFace相似的结果[26]。

本文的贡献总结如下

* •我们建立一个大规模的人脸数据集并将其公之于众，这将消除LFW评估的混乱，并使这些方法具有可比性;
* •我们提出了一种半自动管道来构建来自互联网的大规模人脸数据集，这将吸引更多的研究人员建立新的人脸数据集或扩大现有的人脸数据集;
* •我们培训高性能基线深CNN，用于野外人脸识别。 使用单一网络时，我们的网络性能优于DeepFace [26]和DeepID2 [24]。

# Related Work

数据和算法是模式识别的两个基本组成部分。 随着深度学习在人脸识别中的成功应用，数据集收集滞后于算法。 在本节中，我们将回顾一些流行的面部数据集和表示学习方法。

## Face Dataset

早期的人脸数据集几乎是在受控环境下收集的，如PIE [22]，FERET [21]等。通过许多研究人员的努力，我们可以在这些理想的数据集上获得非常高的性能。但我们发现从这些数据集中学习的模型很难在实际应用中推广到新环境。为了改进人脸识别方法的普遍化，社区的利益逐渐从受控环境转移到不受控制的环境，即野外的人脸识别。然后，里程碑数据集LFW [7]包括5749个主题，于2007年诞生。

与之前的数据集相比，LFW的最大区别在于图像是从Internet上爬行而不是在多个预定义环境下获取的。因此，LFW在姿势，光照，表现，分辨率，成像装置方面具有更多变化，并且这些因素以随机方式组合在一起。 2009年，基于名称列表的LFW [11]收集了另一个名为PubFig的优秀数据集。尽管PubFig仅包含200个主题，但每个主题的图像数量远远超过LFW，并且它提供了73个属性来描述面部图像。 YTF [28]是另一个基于LFW名称列表的数据集，但它是为基于视频的人脸识别而创建的。 YTF中的所有视频都是从YouTube下载的。由于YouTube上的视频压缩率非常高，因此脸部快照的质量低于LFW。

CACD [3]是2014年收集的用于跨年龄人脸识别的大型数据集，其中包括2,000个科目和163,446个图像。 CACD的规模足以训练深度模型，但数据集包含大量噪声和不正确的身份标签。原因是图像被Google Image搜索引擎抓取，只有一小部分（200个主题）通过手动检查。

除上述可公开访问的数据集外，还有三个大型私有数据集：Facebook的SFC [26]，CUHK的CelebFaces [24]和MSRA的WDRef [4]。其中，证监会规模最大，包括4000多个科目，每个科目平均有1000个图像。使用SFC，[26]成功地学习了一种有效的面部表征，可以有效地应对野外变化。虽然CelebFaces和WDRef的规模相对小于SFC，但它们也是开发高性能算法的良好资源。 LFW目前最先进的准确度是通过CelebFaces培训获得的。遗憾的是，这三个优秀的数据集尚未公开，因此本文收集CASIA-WebFace来填补这一空白。

2.2。面部表征学习

第一种流行的人脸识别方法是1991年提出的Eigenface [27]。现在我们可以看到Eigenface是一个具有一个线性层的模型。 Fisherface [1]或LDA也是单层线性模型。在接下来的很长一段时间里，研究人员主要关注如何根据一些成本函数来解决线性层的参数，例如重建误差和分类误差。由于LDA容易出现SSS问题（小样本量），因此LDA的解决方案也得到了很多关注。

然后，出现了各种基于局部特征的方法，并结合上述线性模型自然使用它们，如Gabor + LDA [18]，LBP + LDA [15]等。虽然本地滤波器的参数是手工制作的，但我们可以粗略地将这些方法看作双层模型。第一层通常以局部和非线性方式应用于输入图像，例如Gabor幅度和LBP编码。大量论文表明，“局部非线性+全连通线性”架构肯定比“全连接线性”架构更好。 [13]中提出的Gabor + LBP + LDA甚至是一个三层模型，通过仔细调整获得了良好的性能。

众所周知，文献中很少报道基于手工制作的滤镜的更深层次的模型（#layers> 3）。因为每层的过滤器（或参数）通常是手工独立设计的，并且人类观察难以处理层之间的动态。因此，从数据中学习所有层的参数是最好的方法。随着CNN在图像分类中的成功应用，它迅速成为人脸识别领域的主流。在CNN变得流行之前，许多好的滤波器学习方法[19,14]被提出来学习双层模型的参数，但它们很难推广到深层架构。根据目前的趋势，本文使用深度CNN来学习大规模数据集的人脸表示。

# CASIA-WebFace Dataset

## Name and Image Collection

用于人脸识别的数据集只需要两种数据：人脸图像和身份。从Internet上随机抓取面部图像并对其进行注释几乎是不可能完成的任务。 IMDb是一个结构良好的网站，包含名人的丰富信息，如姓名，性别，生日和照片。我们首先在网站上搜索1940年至2014年间出生的名人，然后抓取他们的名字。

每个名人在网站上都有一个独立的页面。示例页面如图1所示，其中我们只关注“名称”，“主要照片”和“照片库”内容。忽视名人没有“主要照片”，我们共获得38,423个主题和903,304个图像。然后，所有图像都由多视图面部检测器处理，844,126个图像保留在数据集中，并检测到1,556,352个面部。因为许多图像同时出现在serival名人的“照片库”中，所以图像和面部的实际数量小于上面的数字。

## Face Annotation

当前状态的数据集不能用于训练，我们需要在每个图像中注释面部的身份。 “主照片”通常仅包含相应名人的单个面部，但“照片库”中的大多数照片包含属于其他名人的多个面部。 我们的任务是为每个面部分配一个身份，并根据他们的身份将面部分成几组。

main photo name



photo gallery

Figure 1. A sample page of David Fincher on IMDb. The “main photo” is used as initial seed and the 58 photos in the “photo gallery” need to be annotated.



Figure 2. Two sample photos of Ben Affleck containing multiple faces. The name tags corresponded to the photo are shown at the left-bottom of photo. The left photo contains 3 faces and is corresponded to 3 names, but 2 faces are not detected (white rectangles). The right photo contains 3 faces but is only corresponded to 2 names. The woman in the right figure is not annotated (yellow rectangle).

浏览每个名人的“照片库”，我们发现每张照片都有几个名称标签注释。 名称标签可以减少面部身份对应的搜索空间，简化我们的注释任务。 图2中显示了两个照片样本页面，其中说明了照片页面中的两种噪声：未命中检测和未命中注释。

通过现有的面部识别引擎对所有面部进行聚类是处理这种大规模任务的自然方式。 一般聚类方法需要首先计算所有样本的相似性（或距离）矩阵，但矩阵太大而无法加载到内存中。 为了有效地注释大型面部，我们提出了一种通过同时使用名称标签和面部相似性的三步方法。 所提出的方法可以在普通PC上运行并获得良好的聚类结果。 我们的标签相似性聚类方法的步骤如下。

1.通过预训练的人脸识别引擎[29]提取每个人脸的特征模板;

2.使用每个名人的“主要照片”作为其种子。

3.使用包含1张脸部的图像来增强每个名人的播种图像。

4.对于“照片库”中的剩余图像，找到受相似性和名称标签约束的面部和名人之间的对应关系。

5.从图像裁剪面部并保存到每个名人的独立文件夹中。 手动检查数据集并删除错误的分组面部图像。

在完成聚类之后，我们移除具有少于15个面部图像的对象。 为了使此数据集与LFW兼容，我们根据CASIA-WebFace和LFW中名称之间的编辑距离检查重复的主题。 在CASIA-WebFace和LFW之间发现了1043个具有相同名称的科目，这些科目从CASIA-WebFace中删除。 现在，我们可以将CASIAWebFace视为LFW的独立训练集。 通过结合CASIA-WebFace和LFW，我们获得了大规模人脸识别的新基准。

在清理之后，CASIA-WebFace最终拥有10,575个主题和494,414个面部图像。 由于数据集的规模太大，我们无法保证所有面都被正确检测和注释。 少量未命中分类样本不会影响训练过程，并且可能能够提高模型的稳健性。 实验结果将说明数据集的质量。

## 数据集统计

CASIA-WebFace数据集的统计数据如表1所示。除了Facebook的SFC数据集，CASIA-WebFace的规模最大。 对于用户的隐私问题，也许证监会永远不会向研究界开放。 微软WDRef数据集的功能从2012年开始公开，但对于高级研究来说它是不灵活的。 在表格中列出的数据集中，CASIA-IMDb + LFW是野外大规模人脸识别的最佳组合。 如果您认为LFW的准确性已经被当前最先进的方法所饱和[24]。 BLUFR [17]是一个更具挑战性的报告结果的协议。

# 4.学习深度表达

为了说明所提出的数据集的优点，我们通过深度卷积网络以及许多最新技巧从该数据集中学习有效表示。

## 4.1。 卷积网络

基线深度卷积网络是通过结合最近成功网络的许多技巧构建的，包括非常深的架构[23]，低维表示和多重损失函数[24]。小滤波器和非常深的架构可以减少参数的数量并增强网络的非线性。低维表示符合面部图像通常位于低维流形上的假设，并且低维约束可以降低网络的复杂性。在[24]中分析了梳理识别和验证丢失函数，它们可以比Softmax学习更多的判别表示。

输入层的尺寸是100×100×1通道，即灰度图像。所提出的网络包括10个卷积层，5个池化层和1个完全连接的层，其详细架构如表2所示。网络中所有滤波器的大小为3×3。前四个池化层使用max运算符，最后一个池化层是平均值。受到我们GPU的计算能力的限制，该架构不是最佳的，而是根据我们的经验确定的。仍有一些改进空间。

在[23]和[25]中独立提出了小滤波器和非常深的架构。 [23]通过19层网络实现了ImageNet 2014挑战的高性能。同时，[25]通过22层网络获得了比[23]略好的结果。本文结合了这两篇论文的技巧。我们使用多个小滤波器来近似大滤波器并移除冗余的完全连接层以减少参数的数量。最后，我们的网络在所有10个卷积层中使用3×3滤波器，并且只有1个完全连接的层。

Pool5层用作面部表示，其尺寸等于Conv52,320的通道数。为了区分训练集（10575）中的大量主体，这种低维表示应该完全从面部图像中提取判别信息。。 与[24]相同，将Softmax（识别）和对比（验证）成本结合起来构建目标函数。 除了Conv52之外，在所有卷积层之后使用ReLU神经元。 因为Conv52通过平均值组合以生成低维面部表示，所以它们应该是密集且紧凑的。 ReLU易于产生稀疏矢量，因此将其应用于面部表示会降低性能。

在训练阶段，Pool5用作对比成本函数的输入。 并且Fc6用作Softmax成本函数的输入。 因为Fc6的参数数量非常大，即320×10575，所以我们将丢失比设定为0.4以使Fc6正常化。 两个成本函数的重要性由权重α平衡。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | Dataset | #Subjects | #Images | Availability | | LFW [7] | 5,749 | 13,233 | Public | | WDRef [4] | 2,995 | 99,773 | Public (feature only) | | CelebFaces [24] | 10,177 | 202,599 | Private | | SFC [26] | 4030 | 4,400,000 | Private | | CACD [3] | 2,000 | 163,446 | Public (partial annotated) | | CASIA-WebFace | 10,575 | 494,414 | Public |   Table 1. The information of CASIA-WebFace and comparison to other large scale face datasets.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Name | Type | Filter Size  /Stride | Output size | Depth | #Params | | Conv11 | convolution | 3×3 / 1 | 100×100×32 | 1 | 0.28K | | Conv12 | convolution | 3×3 / 1 | 100×100×64 | 1 | 18K | | Pool1 | max pooling | 2×2 / 2 | 50×50×64 | 0 |  | | Conv21 | convolution | 3×3 / 1 | 50×50×64 | 1 | 36K | | Conv22 | convolution | 3×3 / 1 | 50×50×128 | 1 | 72K | | Pool2 | max pooling | 2×2 / 2 | 25×25×128 |  |  | | Conv31 | convolution | 3×3 / 1 | 25×25×96 | 1 | 108K | | Conv32 | convolution | 3×3 / 1 | 25×25×192 | 1 | 162K | | Pool3 | max pooling | 2×2 / 2 | 13×13×192 | 0 |  | | Conv41 | convolution | 3×3 / 1 | 13×13×128 | 1 | 216K | | Conv42 | convolution | 3×3 / 1 | 13×13×256 | 1 | 288K | | Pool4 | max pooling | 2×2 / 2 | 7×7×256 | 0 |  | | Conv51 | convolution | 3×3 / 1 | 7×7×160 | 1 | 360K | | Conv52 | convolution | 3×3 / 1 | 7×7×320 | 1 | 450K | | Pool5 | avg pooling | 7×7 / 1 | 1×1×320 |  |  | | Dropout | dropout (40%) |  | 1×1×320 | 0 |  | | Fc6 | fully connection |  | 10575 | 1 | 3305K | | Cost1 | softmax |  | 10575 | 0 |  | | Cost2 | contrastive |  | 1 | 0 |  | | Total |  |  |  | 11 | 5015K |   Table 2. The architecture of the proposed baseline convolutional network. |

## 4.1。 培训方法

在输入网络之前，所有人脸图像都会转换为灰度，并根据两个标记归一化为100×100（参见图4）。与最常用的眼睛中心相比，这里所选择的两个界标之间的距离对于偏航角的姿势变化是相对不变的。归一化后，两点之间的距离为25像素。由于人脸具有接近对称的结构，我们通过镜像操作使训练集加倍，这可以使表示对姿势变化更加鲁棒。

由于存在如此大量的数据，当前网络不太可能过度拟合，因此我们将所有卷积层的权重衰减设置为0，并将完全连接层的权重衰减设置为5e-4。学习率最初设定为1e-2，逐渐减少到1e-5。由于Softmax的收敛速度快于对比成本函数，因此权重α首先设置为小值3.2e-4，并逐渐增加到6.4e-3。

开源实现cuda-convnet [9]用于训练我们的网络。对于Softmax成本，我们只需要输入面部图像及其标签，但是对于对比成本，我们需要通过从训练集中采样来生成面部对。为了减少内存和磁盘空间的消耗，我们只是在每个批次中在线对正面和负面对进行采样。不包括批次中的面对。如何有效地生成完整的面部对将留待将来的工作。

# Experiments

CASIA-WebFace始终用于训练深层网络。 如上一节所述，CASIA-WebFace中的所有面部图像都通过面部检测，面部标记和对齐处理。 由于面部对称，我们镜像493,456个检测到的面以增加数据集。 最后，我们有986,912个训练样本。 整个过程是完全自动的，错误对齐的面部保持不变。 此数据集中仍有少量未命中检测。 如果您有比我们更好的面部检测器，您的训练集可能比我们的稍微大一些。

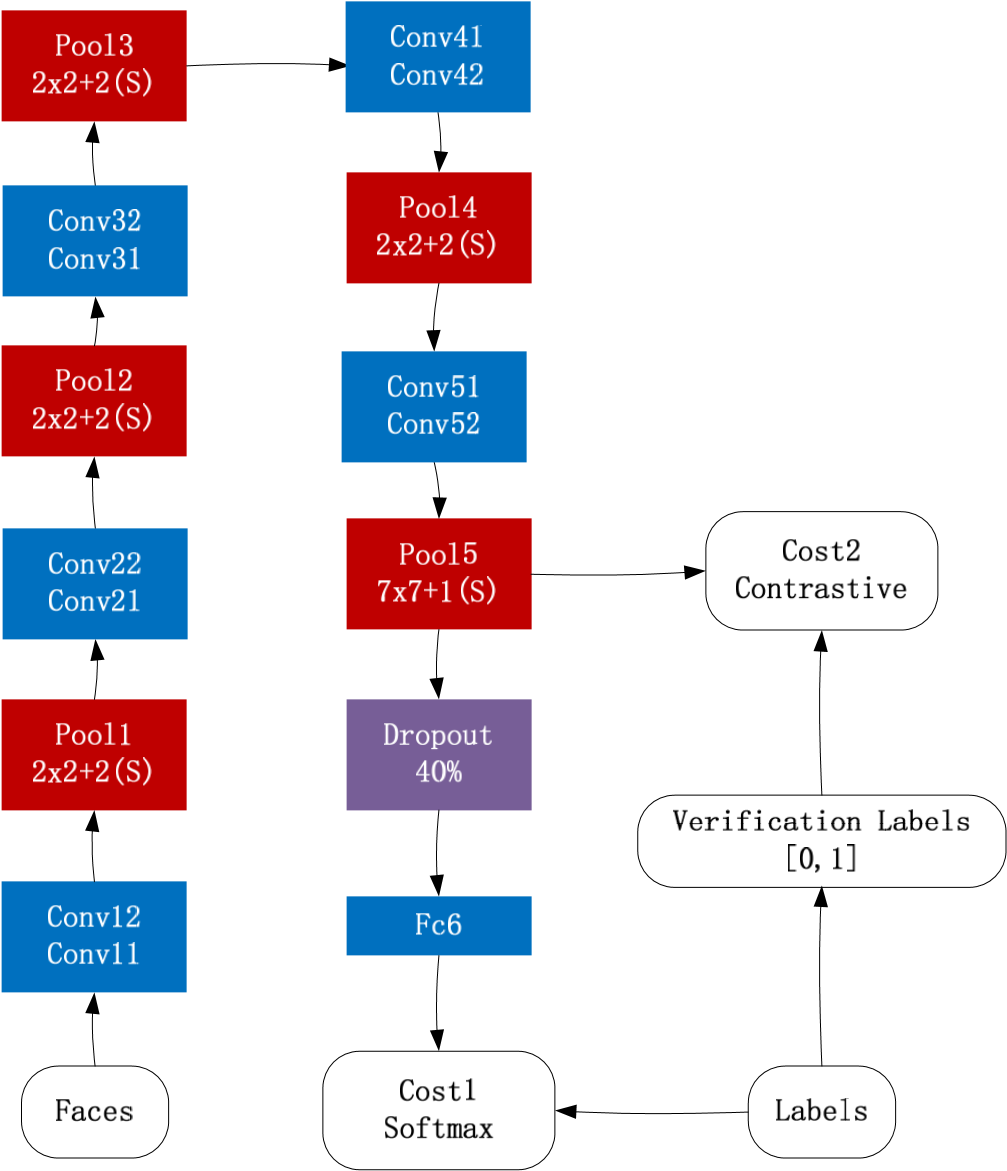


Figure 3. The proposed baseline convolutional network with many recent tricks.

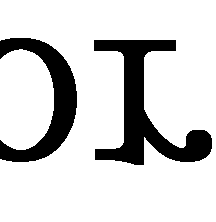
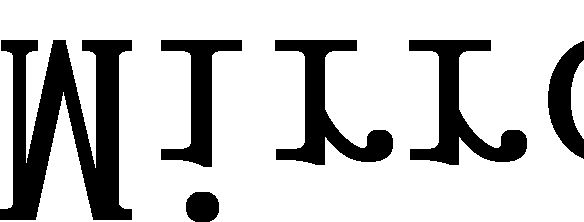
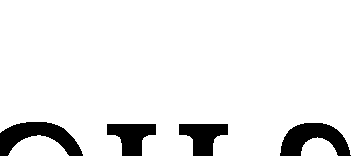
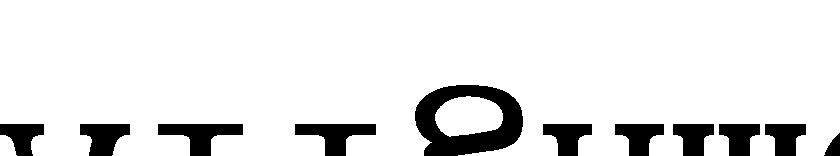
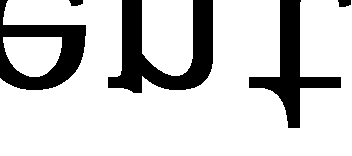
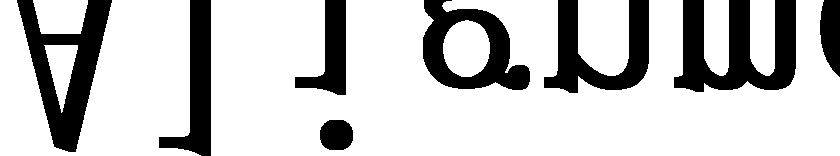


Figure 4. Face image alignment and augmentation. The read circles on the face are two selected landmarks for similarity transformation.

LFW和YTF是野外人脸识别中最受欢迎和最具挑战性的两个数据集。 因为它们与提议的CASIA-WebFace不重叠，所以报告LFW和YTF的性能是非常合理的。 除此之外，训练有素的深层网络也根据更具挑战性和实用性的协议BLUFR进行评估，该协议能够真实地反映人脸识别的性能。...

|  |  |  |  |
| --- | --- | --- | --- |
| Method | #Net | Accuracy±SE | Protocol |
| DeepFace | 1 | 95*.*92±0*.*29% | unsupervised |
| DeepFace | 1 | 97*.*00±0*.*28% | restricted |
| DeepFace | 3 | 97*.*15±0*.*27% | restricted |
| DeepFace | 7 | 97*.*35±0*.*25% | unrestricted |
| DeepID2 | 1 | 95*.*43% | unrestricted |
| DeepID2 | 2 | 97*.*28% | unrestricted |
| DeepID2 | 4 | 97*.*75% | unrestricted |
| DeepID2 | 25 | 98*.*97% | unrestricted |
| Ours A | 1 | 96*.*13±0*.*30% | unsupervised |
| Ours B | 1 | 96*.*30±0*.*35% | unsupervised |
| Ours C | 1 | 97*.*30±0*.*31% | unsupervised [[1]](#footnote-1) |
| Ours D | 1 | 96*.*33±0*.*42% | unsupervised |
| Ours E | 1 | 97*.*73±0*.*31% | unrestricted |

Table 3. The performance of our baseline deep networks and compared methods on LFW View2.

## Results on LFW

LFW包括5,749个主题和13,233个面部图像。 性能报告有三种主要协议：无监督，受限和不受限制的协议。 无监督协议用于评估人脸表示的基线性能，其他两个协议通常用于评估度量学习或整个方法的性能。 对于所有协议，测试集是固定的，其中包括10个分割中的6000个面对。 应报告平均值的平均准确度和标准误差。

### 标准协议

LFW中的所有图像都由与CASIA-WebFace相同的管道处理，并标准化为100×100。 所有面部的表示由在CASIA-WebFace（缩写为DR）上训练的深度网络提取。 首先，我们直接评估表示的基本性能。 然后，我们测试无监督和监督学习对基本表示的影响。 进行以下实验：

* •答：DR +余弦;
* •B：CASIA上的DR + PCA-WebFace + Cosine;
* •C：DR +联合贝叶斯CASIA-WebFace;
* •D：LFW训练集+余弦训练中的DR + PCA;
* •E：LFW训练集上的DR + Joint Bayse。

根据LFW的协议，超参数在CASIA-WebFace或LFW的View1上进行调整，例如PCA的维数和Joint Bayes的正则化因子。精度在LFW的View2上进行评估，并列于表3中。还提供了DeepFace和DeepID2的其他最新结果用于比较。从我们的6个实验结果中，我们可以得出4个结论：

1.基本表示具有良好的性能;

2.对LFW训练集进行微调可以略微提高性能，例如B→D，C→E;

3.基于基本表示，联合贝叶斯可以略微提高性能，例如B→C，D→E。

通过在无监督设置中检查结果，我们可以看到我们的基本表示优于DeepFace，96.13％对95.92％。通过PCA调整LFW后，精度96.33％略有提高。由于Joint Bayes不能直接处理成对样本，因此我们不会通过限制协议进行实验。当使用不受限制的协议时，我们的单网络方案E达到97.73％，这比DeepFace的7网络集合97.35％更好，并且与DeepID2的4网络集合97.75％相当。

我们的网络优势受益于深层架构。在其他方面，我们的方法和数据集仍然不如DeepFace：1）我们只是通过2D相似变换对齐人脸图像，这不如DeepFace的3D对齐; 2）DeepFace，SFC的训练集规模比CASIA-WebFace大10倍。受GPU计算资源的限制，我们不会继续培训更深入的网络或训练网络集合来改善这里的性能。在发布数据集CASIAWebFace之后，我们相信整个研究社区可以更快地刷新记录。

### BLUFR协议

LFW的测试集仅包括6000个面对，其中一半是真实的，另一半是冒名顶替者。 如[17]中所讨论的，这种负面对的比例不足以评估低FAR的性能。 因此，[17]开发了一种新的基准协议，称为BLUFR，以充分利用LFW中的所有13,233个人脸图像。 BLUFR包含验证和开放式识别方案，重点是低FAR。 有10个实验试验，每个试验包含约156,915个真实匹配分数和46,960,863个冒名顶替匹配分数，用于绩效评估。

LFW中面部的表示以与先前实验相同的方式提取。 然后可以通过标准基准工具包[16]计算结果。 为简单起见，我们只报告方案E和F的结果。

|  |  |  |
| --- | --- | --- |
| Method | VR@FAR=0.1% | DIR@FAR=1%  Rank=1 |
| HD-LBP + JB | 41*.*66% | 18*.*07% |
| HD-LBP + LDA | 36*.*12% | 14*.*94% |
| Ours E | 80*.*26% | 28*.*90% |

Table 4. The performance of our baseline deep network and compared methods on LFW under BLUFR protocol.

表4中列出了我们的方法和比较方法的VR（验证率）和DIR（检测和鉴定率）。表中的数字是在10次试验的（μ-σ）中测量的，其中μ是平均准确度和σ是标准偏差。

[17]刚刚报道了BLUFR协议下一些传统浅（但宽）模型的性能。报道最好的方法是HD-LBP + JB（高维LBP +关节贝叶斯），其结果是VR = 41.66％（FAR = 0.1％）。如表4所示，我们的深度网络显着优于基于HD-LBP的方法。深模型对宽模型的优越性已在以前的工作中得到说明，并且本文的结论再次得到验证。

我们发现表4中的所有数字明显低于表3中的数字，尤其是DIR（FAR = 1％且Rank = 1）。由于DIR是反映人脸监视（或观察列表）系统性能的重要指标，我们认为人脸识别算法仍然存在很大差距，无法满足监控应用的要求。

## Results on YTF

* 为了测试我们网络的泛化能力，我们在视频人脸数据集YouTube Faces（YTF）上对其进行评估。 由于运动模糊和高压缩比，YTF中的图像质量比网络照片差。 对于YTF中的每个视频，我们随机选择15帧并通过深层网络（DR）提取其表示。 在训练阶段，15帧被视为具有相同身份的15个样本。 在测试阶段，视频对的相似度得分是15×15 = 225帧对的平均值。 以下实验在无人监督和监督的环境中进行：
* •答：DR +余弦;
* •D：YTF训练集+余弦训练中的DR + PCA;
* •E：YTF训练集上的DR +联合贝叶斯。

DeepFace在YTF上保持最佳结果，并且比其他方法更好，因此，我们只与DeepFace进行比较。 通过余弦函数直接匹配，基本表示在YTF上达到88.00％的准确度。 通过PCA在YTF上转换表示，精度显着提高到90.60％。 当进一步调整Joint Bayes的表示时，我们的方法略微优于DeepFace。...

|  |  |  |  |
| --- | --- | --- | --- |
| Method | #Net | Accuracy | Protocol |
| DeepFace | 1 | 91*.*4±1*.*1% | supervised |
| Ours A | 1 | 88*.*00±1*.*50% | unsupervised |
| Ours D | 1 | 90*.*60±1*.*24% | unsupervised |
| Ours E | 1 | 92*.*24±1*.*28% | supervised |

Table 5. The performance of our methods and DeepFace on Youtube Faces (YTF).

# Conclusion

这项工作收集了一个来自互联网的大规模人脸数据集，并将其公之于众。 新数据集与LFW不重叠，可与LFW结合用于大规模人脸识别研究。 这种组合可以使LFW的评估协议标准化并推进可重复的研究。 另一方面，统一的培训和测试集可以使各种方法具有可比性。 这项工作还描述了11层卷积网络的数据集构建和人脸表示学习的整个过程。 参考本文提出的管道，任何人都可以轻松培养出高性能的人脸识别引擎。 未来的工作将在三个方向完成：1）通过使用商业图像搜索引擎来增加数据集; 2）开发更有效的注释工具和算法; 3）探索训练单个网络接近大型深度网络集成性能的新方法。

# Acknowledgements

This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No.KGZD-EW-102-2, and AuthenMetric R&D Funds. The Tesla K40 used for this research was donated by the NVIDIA Corporation.

# References

1. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman.

“Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection”. In *Proceedings of the European Conference on Computer Vision*, pages 45–58, 1996. 2

1. Y. Bengio. “Learning deep architectures for AI”. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009. 1
2. B.-C. Chen, C.-S. Chen, and W. H. Hsu. “Cross-age reference coding for age-invariant face recognition and retrieval”. In *Computer Vision–ECCV 2014*, pages 768–783. Springer,

2014. 2, 5

1. D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. “Bayesian face revisited: A joint formulation”. In *Computer Vision– ECCV 2012*, pages 566–579. Springer, 2012. 2, 5
2. D. Chen, X. Cao, F. Wen, and J. Sun. “Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification”. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032, 2013. 1
3. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. “Improving neural networks by preventing co-adaptation of feature detectors”. *arXiv preprint arXiv:1207.0580*, 2012. 2
4. G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. “Labeled faces in the wild: A database for studying face recognition in unconstrained environments”. 2007. 1, 2, 5
5. Y. Jia. Caffe. [http://caffe.berkeleyvision. org/.](http://caffe.berkeleyvision.org/) 1
6. A. Krizhevsky. cuda-convnet. [http://code.google. com/p/cuda-convnet/.](http://code.google.com/p/cuda-convnet/) 1, 5
7. A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1
8. N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar.

Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009. 2

1. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradientbased learning applied to document recognition”. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1
2. Z. Lei, S. Liao, M. Pietikainen, and S. Z. Li. “Face recognition by exploring information jointly in space, scale and orientation”. *Image Processing, IEEE Transactions on*, 20(1):247–256, 2011. 3
3. Z. Lei, M. Pietikainen, and S. Li. “Learning discriminant face descriptor”. 2013. 3
4. S. Z. Li, R. Chu, S. Liao, and L. Zhang. “Illumination invariant face recognition using near-infrared images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, April 2007. 3
5. S. Liao. “Benchmark of large-scale unconstrained face recognition”. [http://www.cbsr.ia.ac.cn/users/ scliao/projects/blufr/.](http://www.cbsr.ia.ac.cn/users/scliao/projects/blufr/) 7
6. S. Liao, Z. Lei, D. Yi, and S. Z. Li. “A benchmark study of large-scale unconstrained face recognition”. In *IAPR/IEEE International Joint Conference on Biometrics, Clearwater, Florida, USA*, 2014. 2, 4, 7
7. C. Liu and H. Wechsler. “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition”. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002. 3
8. D. Maturana, D. Mery, and A. Soto. “Face recognition with decision tree-based local binary patterns”. In *Computer Vision–ACCV 2010*, pages 618–629. Springer, 2011. 3
9. V. Nair and G. E. Hinton. “Rectified linear units improve restricted boltzmann machines”. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010. 2
10. P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. “The FERET evaluation methodology for face-recognition algorithms”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000. 2
11. T. Sim, S. Baker, and M. Bsat. “The cmu pose, illumination, and expression (PIE) database”. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE, 2002. 2
12. K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4
13. Y. Sun, X. Wang, and X. Tang. “Deep learning face representation by joint identification-verification”. *arXiv preprint arXiv:1406.4773*, 2014. 1, 2, 4, 5
14. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going deeper with convolutions”. *arXiv preprint arXiv:1409.4842*, 2014. 4
15. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. “Deepface: Closing the gap to human-level performance in face verification”. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 1, 2, 5
16. M. A. Turk and A. P. Pentland. “Eigenfaces for recognition”. *Journal of Cognitive Neuroscience*, 3(1):71–86, March 1991. 2
17. L. Wolf, T. Hassner, and I. Maoz. “Face recognition in unconstrained videos with matched background similarity”. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011. 2
18. D. Yi, Z. Lei, and S. Z. Li. “Towards pose robust face recognition”. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3539–3545, 2013.

4

1. The term “unsupervised” means that the model is not trained on LFW in supervised way. [↑](#footnote-ref-1)