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

DeepFace系统概述

主要工作：3D人脸校正+9层神经网络

训练数据：

SFC库——4030人，440万图片（800-1200张/人）

额外训练数据：

10万人，300万图片（30张/人）

测试数据：LFW，YTF

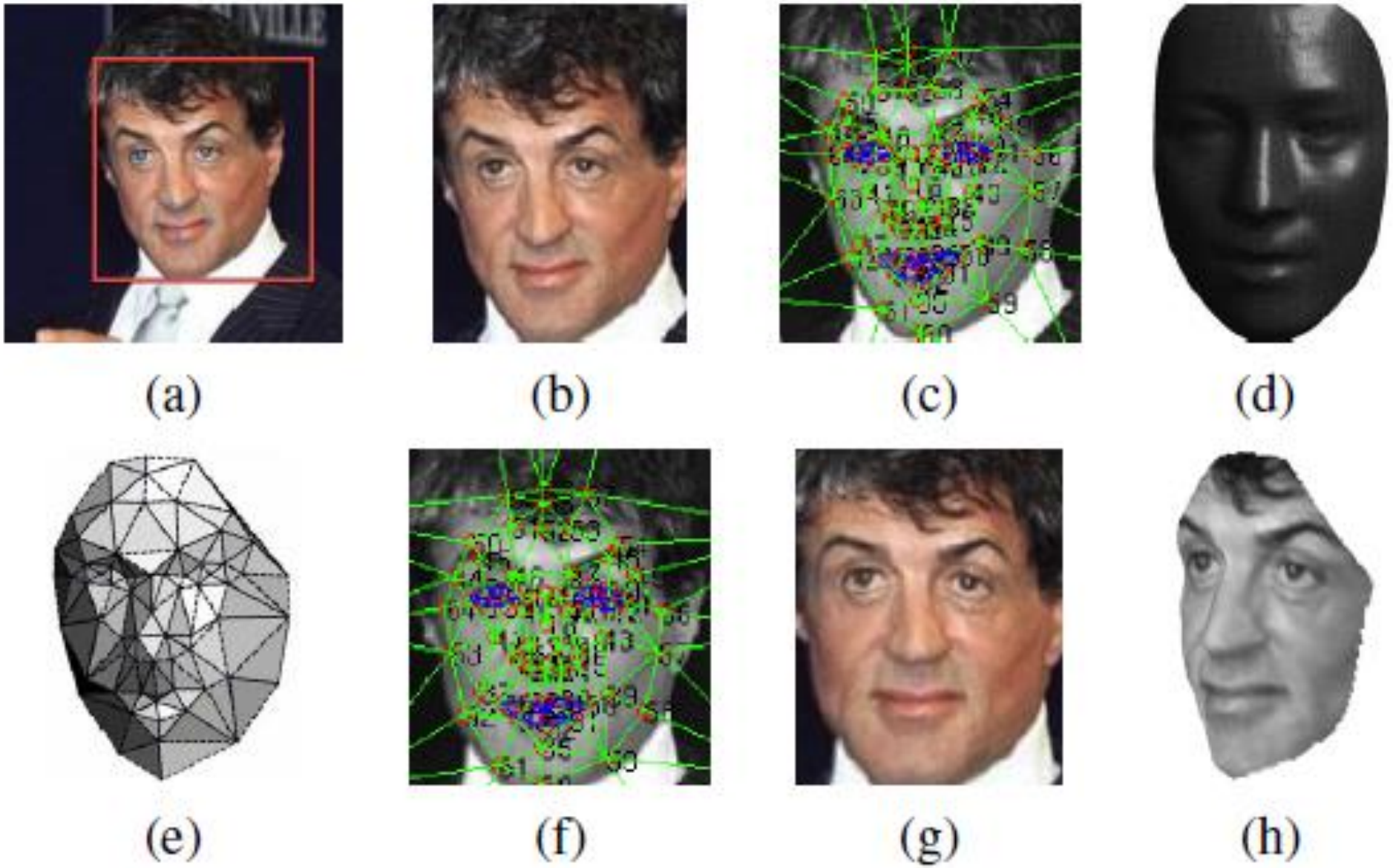
成果：97.35% in LFW，3D校正系统，泛化性

DeepFace 优势

深度学习：特别适合解决大样本训练问题，传统的机器学习方法在使用样本训练时存在容量上限的问题。

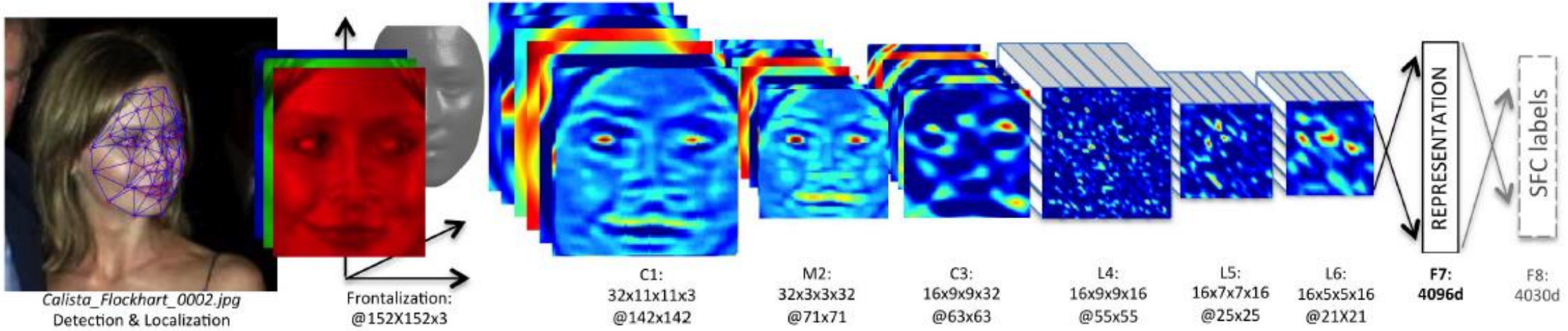
计算资源：大规模的计算资源（大量的CPU或者GPU）更容易获得

3D校正：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像素信息中进行学习。



（a）检测到的人脸图；（b）2D校正，相似变换；（c）2D校正人脸上的67个基准点，及相应的Delaunay三角形；（d）3D形状变换模型；（e）3D-2D成像机；（f）结合3D模型进行分块的仿射变换；（g）最终得到的正面人脸；（h）使用3D模型产生的图片

DNN的结构和训练



输入3通道的人脸，并进行3D校正，归一化到152*152	卷积层C1：32个11*11*3的滤波器	Max-polling：M2滑窗大小为3*3，滑动步长为2	卷积层C3：16个9*9*16的3维卷积核。提取到简单的边缘特征和纹理特征	L4,L5,L6都是局部连接层，在特征图像的每一个位置都训练学习一组不同的滤波器	F7,F8是全连接的：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。

DNN的结构和训练

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。

为了提取到低水平的特征，如简单的边缘特征和纹理特征。

DNN的结构和训练

5，L4,L5,L6都是局部连接层，在特征图像的每一个位置都训练学习一组不同的滤波器。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。

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。

DNN的结构和训练

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o_k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：

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

使得下式（叉熵损失）最小，即是最大化了正确输出类别的概率

$$L = -\log p_k$$

使用神经网络提取人脸特征

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

$$\phi = \{C_1,...,F_7\}, \quad \theta_T = \{x_{2d}, \vec{P}, \vec{r}\}$$

表示图像的姿态等信息。

把特征的元素归一化成0到1（ L2归一化），以此降低特征对光照变化的敏感度。

人脸验证——距离度量方法

提取到的特征的特点：

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。

适合采用加权的 χ^2 距离

$$\chi^2(f_1, f_2) = \sum_i w_i (f_1[i] - f_2[i])^2 / (f_1[i] + f_2[i])$$

Wi由线性SVM学习到

人脸验证——距离度量方法

Siamese network(一个监督的度量学习模型)

对输入的两张图片提取特征，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。

a,计算两个特征之间的绝对差别；

b,接一个全连接层，映射到一个逻辑输出单元（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，也是按相同的方法学习得到。

实验

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，主要是视频段。

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）

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）,用来评估视频级的人脸验证。

图片质量：SFC<LFW，YTF

标注误差：SFC（3%），YTF（100个视频对）。

分类误差的评估

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%

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%，这是因为训练集骤减后，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**。

LFW上的人脸验证

人类在LFW上人脸验证准确度为97.5%（裁剪后的人脸）

DeepFace=3D校正+DNN

1，倘若用2D的校正，准确率为94.3%，完全不校正，准确率为87.9%；

2，不用DNN，用3D校正结合朴素LBP/SVM,达到91.4%。

3, 直接比较归一化之后的特征对的内积。达到了95.92%的准确率（无监督）

4，χ^2距离度量，97%

5，调整不同的输入类型训练到多个DNN，将各网络的结果结合起来，97.15%

LFW上的人脸验证

- 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

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

97.15%

LFW上的人脸验证

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%**。

几种网络的验证性能

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

与其他方案的对比

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.

视频级的人脸验证

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）。

视频级的人脸验证

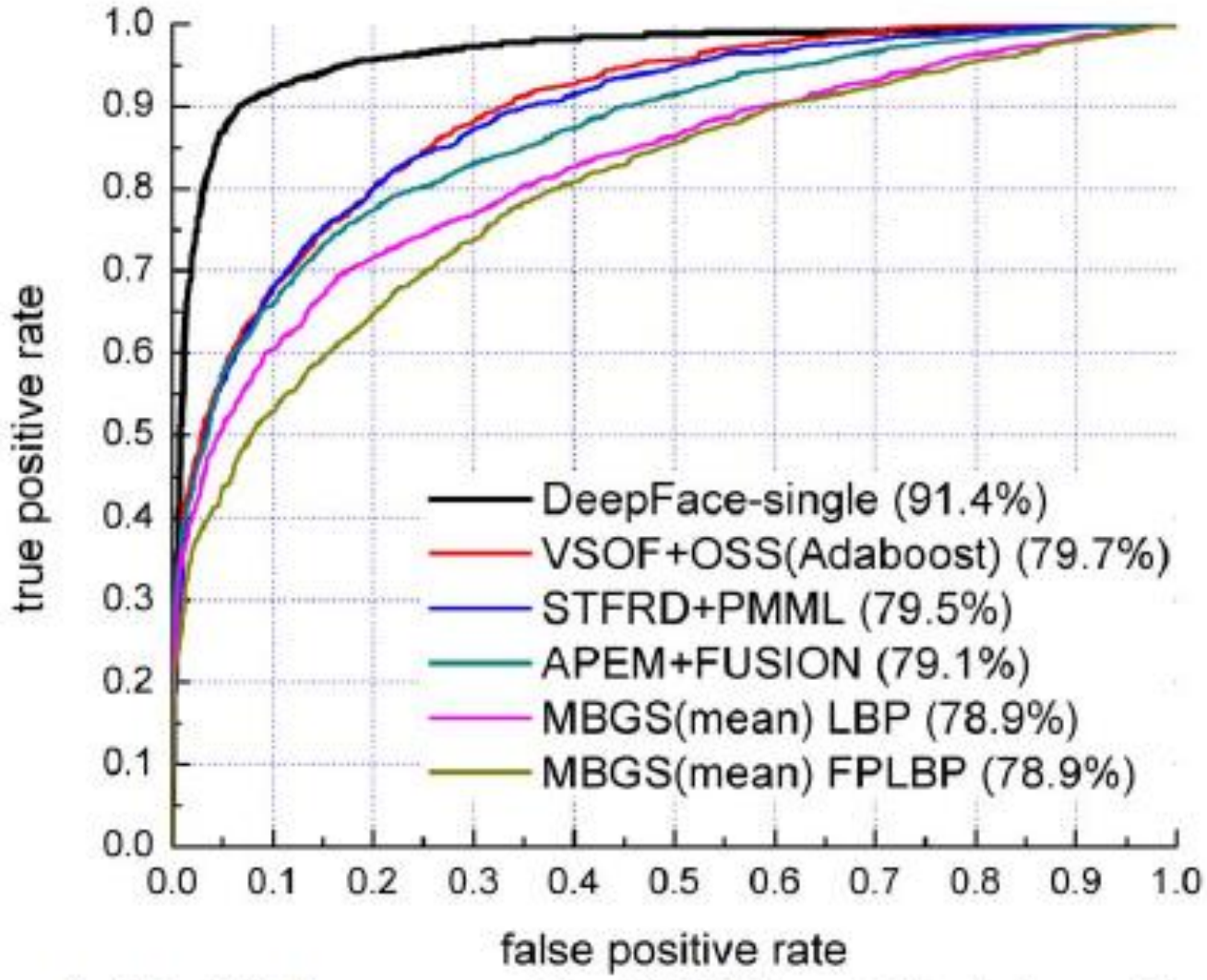


Figure 4. The ROC curves on the *YTF* dataset. Best viewed in color.

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DNN+输出最终分类。