人脸检测与人脸识别

一、概述

1. 基本概念

人脸是个人重要的生物特征，业界很早就对人脸图像处理技术进行了研究。人脸图像处理包括人脸检测、人脸识别、人脸检索等。人脸检测是在输入图像中检测人脸的位置、大小；人脸识别是对人脸图像身份进行确认，人脸识别通常会先对人脸进行检测定位，再进行识别；人脸检索是根据输入的人脸图像，从图像库或视频库中检索包含该人脸的其它图像或视频。

2. 人脸检测与识别的应用

- 实名认证
- 人脸考勤
- 刷脸支付、刷脸检票
- 公共安全：罪犯抓捕、失踪人员寻找

3. 传统人脸检测与人脸识别方法

1) 人脸检测

- 基于知识的人脸检测法。它将典型的人脸形成规则库对人脸进行编码。通常, 通过面部特征之间的关系进行人脸定位。
- 基于模板匹配的人脸检测法。存储几种标准的人脸模式, 用来分别描述整个人脸和面部特征;计算输入图像和存储的模式间的相互关系并用于检测。
- 基于特征的人脸检测法。是在姿态、视角或光照条件改变的情况下找到存在的结构特征, 然后使用这些特征确定人脸。
- 基于外观的人脸检测法。与模板匹配方法相反, 从训练图像集中进行学习从而获得模型(或模板) ,并将这些模型用于检测。

2) 人脸识别

- 几何特征分析法。首先，提取目标的特征，并将所有得到的特征值组合形成一个向量；然后利用某种距离公式进行比较匹配。

- 主成分分析法。提取出图像数据主成分，得到特征向量（特征脸）在进行比对和识别。主成分分析法原理简单，容易编程实现，并且识别效果较好；但该方法易受光照、尺度、旋转等因素影响。
- 弹性匹配法。属于动态模板匹配法的一种，模板可有多种表示方法，该方法受图像的形变影响小，且受光照、面部表情、图像尺寸等因素的干扰较小，不足之处在于识别速度慢。

传统人脸检测、识别在特征提取、精确度、可扩展性方面均有诸多不足，进入深度学习时代后，逐渐被深度学习技术所取代。

二、人脸数据集介绍

1. SFC数据集

Social Face Classification（社交人脸分类，简称SFC）数据集是从一个流行的社交网络中收集的人脸数据集，包括440万张经过标记的人脸，来自4030人，每个人有800到1200张人脸图像，其中每个身份的最新5%的脸图像被排除在外进行测试。这是根据图像的时间戳来完成的，以模拟通过老化进行的连续识别。

2. LFW数据集

Labeled Faces in the Wild（经标注的自然条件人脸，简称LFW）包含5749位名人的13323张网络照片，这些照片分为6000对人脸，分为10组。常用作无约束环境中进行人脸验证的基准数据集。

3. YTF数据集

YouTube Faces（YTF）收集了1595个主题的3425个YouTube视频（LFW中名人的子集）。这些视频被分成5000个视频对和10个分割，用于评估视频级别的人脸验证在SFC中，人脸识别是由人来标记的，通常包含大约3%的错误。SFC数据集照片在图像质量、光线和表情方面的变化甚至比LFW和YTF中名人的网络图像更大，后者通常是由专业摄影师而不是智能手机拍摄的。

4. CelebA数据集

Large-scale CelebFaces Attributes (CelebA)数据集是由香港中文大学汤晓鸥教授实验室公布的大型人脸识别数据集。该数据集包含有200K张人脸图片，人脸属性有40多种，主要用于人脸属性的识别。

5. WIDER Face数据集

2015年由香港中文大学发布，包含32203张图像、393703张人脸，在面部的尺寸、姿势、遮挡、表情、妆容和光照上都有很大的变化，自发布后广泛应用于评估性能比传统方法更强大的卷积神经网络。

三、人脸检测

1. MTCNN模型

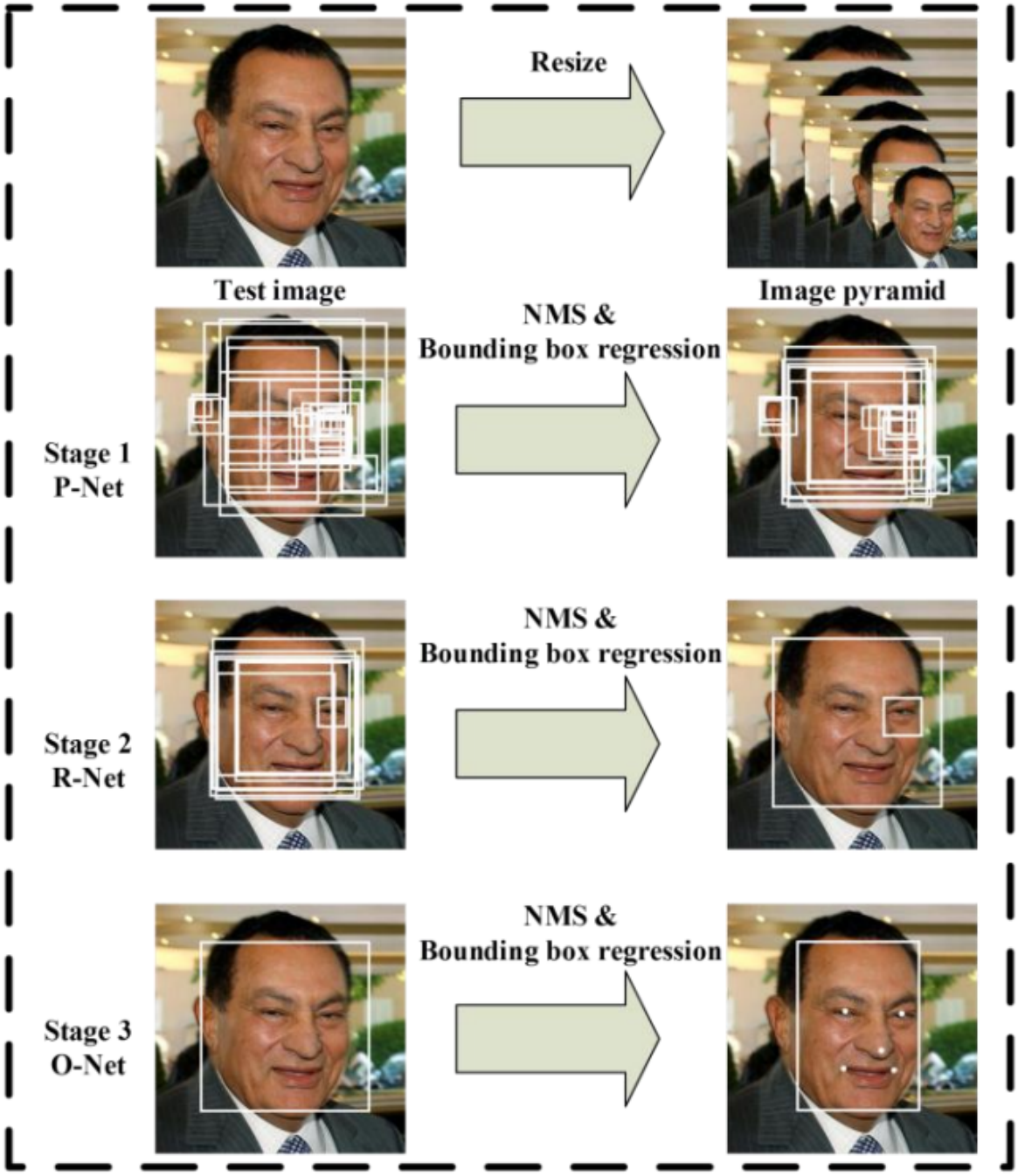
Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks（基于多任务级联卷积网络的联合人脸检测与对准，MTCNN），是一个优秀的人脸检测模型，该模型通过三个阶段精心设计的深度卷积网络，以粗略到精细的方式检测面部位置。

1）步骤

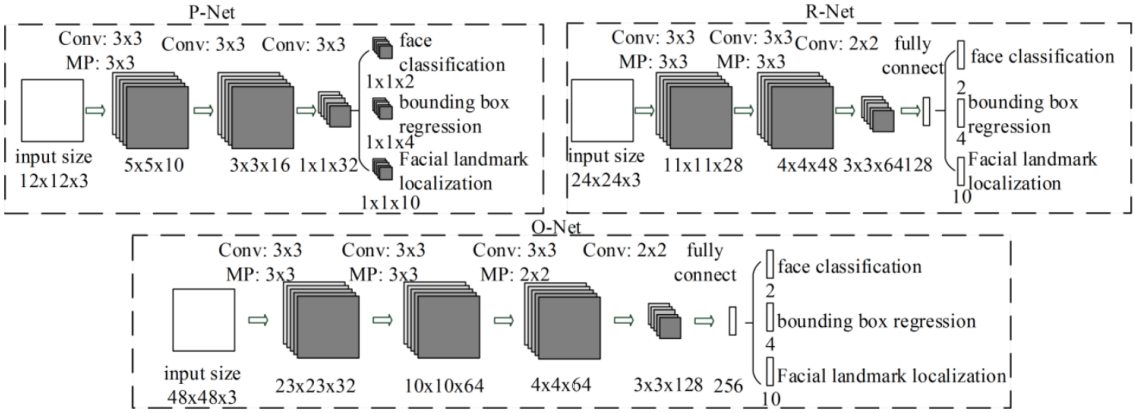
第一阶段：通过浅层CNN快速生成候选窗口。

第二阶段：通过更复杂的CNN拒绝大量非面部窗口来细化窗口。

第三阶段：使用更强大的CNN再次细化结果并输出五个面部标志位置。



2）网络结构



- Proposal Network（P-Net）：提议网络，该完全卷积网络来获得候选面部窗口及其边界框回归向量。 然后基于估计的边界框回归向量校准候选者。 之后，我

们采用非最大抑制（NMS）来合并高度重叠的候选者。

- Refine Network（R-Net）：精炼网络（R-Net），它进一步拒绝大量错误候选者，使用边界框回归执行校准，并进行NMS。
- Output Network（O-Net）：输出网络，这个阶段类似于第二阶段，但在这个阶段，我们的目标是识别更多监督的面部区域。 特别是，该网络将输出五个面部坐标点。

3）训练

MTCNN利用三项任务来训练CNN探测器：

- 面部/非面部分类
- 边界框回归
- 面部标记定位

① 面部分类。学习目标被制定为二类分类问题。 对于每个样本，使用交叉熵损失函数：

$$L_i^{det} = -(y_i^{det} \log(p_i) + (1 - y_i^{det})(1 - \log(p_i))) \quad (1)$$

其中p_i是网络产生的概率，表明样本是网络产生的概率，表明样本x_i是一个人脸。符号 $y_i^{det} \in \{0, 1\}$表示真实标签。

② 边界框回归。对于每个候选窗口，我们预测它与最近的真实值之间的偏移（即边界框的左边，顶部，高度和宽度）。 学习目标被指定为回归问题，我们对每个样本使用欧几里德损失x_i：

$$L_i^{box} = \left\| \hat{y}_i^{box} - y_i^{box} \right\|_2^2 \quad (2)$$

其中$\hat{y}_i^{box}$ 是真实坐标。 有四个坐标，包括左上角，高度和宽度，因此$y_i^{box} \in \mathbb{R}^4$.

③ 面部标记定位。类似于边界框回归任务，面部标记检测被公式化为回归问题，我们最小化欧几里德损失：

$$L_i^{landmark} = \left\| \hat{y}_i^{landmark} - y_i^{landmark} \right\|_2^2 \quad (3)$$

其中 $\hat{y}_i^{landmark}$是从网络获得的面部标记坐标，$y_i^{landmark}$是第i个样本的真实坐标。有五个面部标志，包括左眼，右眼，鼻子，左嘴角和右嘴角，因此$y_i^{landmark} \in \mathbb{R}^10$.

④ 多源训练。由于在每个CNN中使用不同的任务，因此在学习过程中存在不同类型的训练图像，例如面部，非面部和部分对齐的面部。 在这种情况下，不使用一些损失函数（即，等式（1）-（3））。 例如，对于背景区域的样本，仅计算$L_i det$，而另外两个损失设置为0.这可以直接使用样本类型指示符来实现。 然后整体学习目标可以表述为：

$$\min \sum_{i=1}^{N} \sum_{j \in \{det, box, landmark\}} \alpha_j \beta_i^j L_i^j \qquad (4)$$

其中N是训练样本的数量，a_j表示任务重要性。 在P-Net和R-Net中使用$(a_{det} = 1, a_{box} = 0.5, a_{landmark} = 0.5)$，而在O-Net中使用$(a_{det} = 1, a_{box} = 0.5, a_{landmark} = 1)$以获得更准确的面部标记本地化. $\beta_i^j \in \{0, 1\}$是样本类型指示器。

4）结论

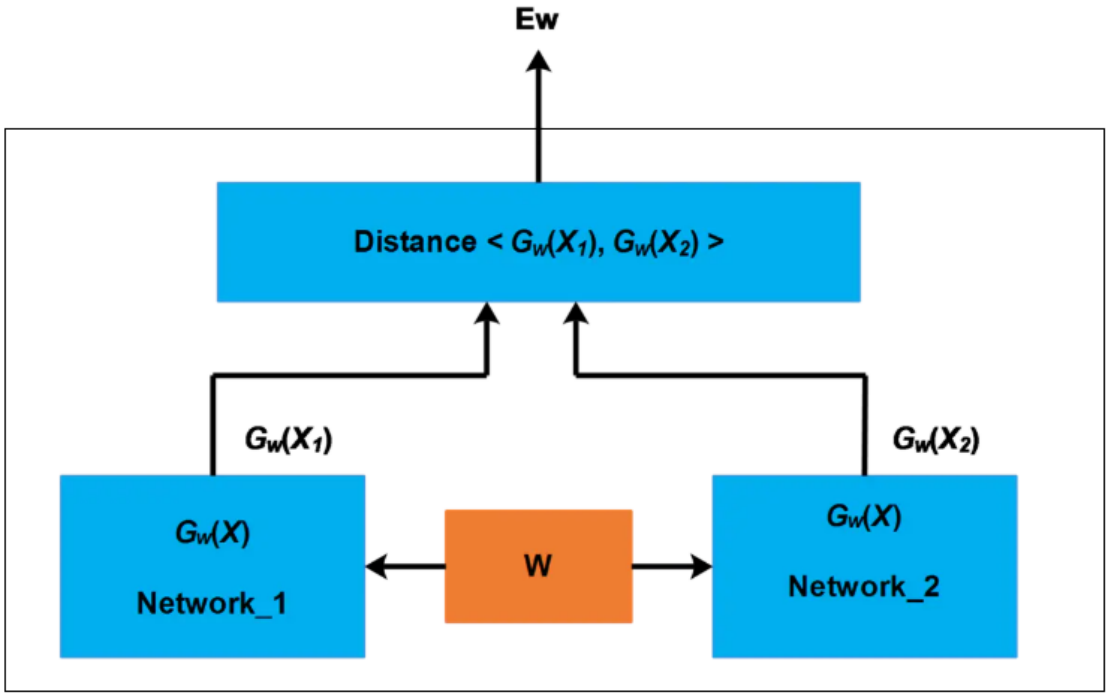
- 查准率：95.04%
- 召回率：85.1%
- 速度：99 FPS

四、单样本学习模型

1. 孪生网络

1）概述

孪生网络（Siamese network）又称“连体网络”，有两个结构相同（两个网络可以是CNN，也可以是LSTM等），且共享权值的子网络。分别接受两个输入x_1, x_2，将其转换为某种向量$G_w(x_1), G_w(x_2)$，再通过某种距离度量的方式计算两个输出向量的距离E_w.如下图所示：



"孪生网络"一词的由来：

> 十九世纪泰国出生了一对连体婴儿，当时的医学技术无法使两人分离出来，于是两人顽强地生活了一生，1829年被英国商人发现，进入马戏团，在全世界各地表演，1839年他们访问美国北卡罗莱那州后来成为“玲玲马戏团”的台柱，最后成为美国公民。1843年4月13日跟英国一对姐妹结婚，恩生了10个小孩，昌生了12个，姐妹吵架时，兄弟就要轮流到每个老婆家住三天。1874年恩因肺病去世，另一位不久也去世，两人均于63岁离开人间。两人的肝至今仍保存在费城的马特博物馆内。从此之后“暹罗双胞胎”（Siamese twins）就成了连体人的代名词，也因为这对双胞胎让全世界都重视到这项特殊疾病。

2）孪生网络的用途

孪生网络主要用于单样本或样本较少的模型训练，用于衡量两个输入数据的相似程度。孪生网络有两个输入（x_1和x_2），将两个输入到两个相同且权重共享的网络中，这两个网络分别将输入映射到新的空间，形成输入在新的空间中的表示。通过Loss的计算，评价两个输入的相似度。例如：

- 词汇或文本的语义相似度分析；
- QA中question和answer的匹配；
- 签名或人脸的比对、验证。

例如，在人脸比对中，如果输入的两幅人脸图像X_1和X_2为同一个人，那么映射到新的空间的两个向量距离足够小；反之，如果X_1和X_2不为同一个人，两个向量距离足够大。

3）损失函数

孪生网络采用对比损失函数（contrastive loss），其表达式如下：

$$(1-Y)\frac{1}{2}(D_W)^2+(Y)\frac{1}{2}\{max(0,m-D_W)\}^2$$

其中，D_W被定义为姐妹孪生网络的输出之间的欧氏距离。D_W欧式距离公式如下：

$$\sqrt{\{G_W(X_1)-G_W(X_2)\}^2}$$

- G_W是其中一个姐妹网络的输出。X_1和X_2是输入数据对;
- Y值为1或0。如果模型预测输入是相似的，那么Y的值为0，否则Y为1。当Y=0时（输入是相似的），函数的值为前半部分；当Y=1时（输入不相似），函数的值为后半部分；
- max（ ）是表示0和$m-D_w$之间较大值的函数;
- m是大于0的间隔值（margin value）。有一个边际价值表示超出该边际价值的不同对不会造成损失。这是有道理的，因为你只希望基于实际不相似对来优化网络，但网络认为是相当相似的。

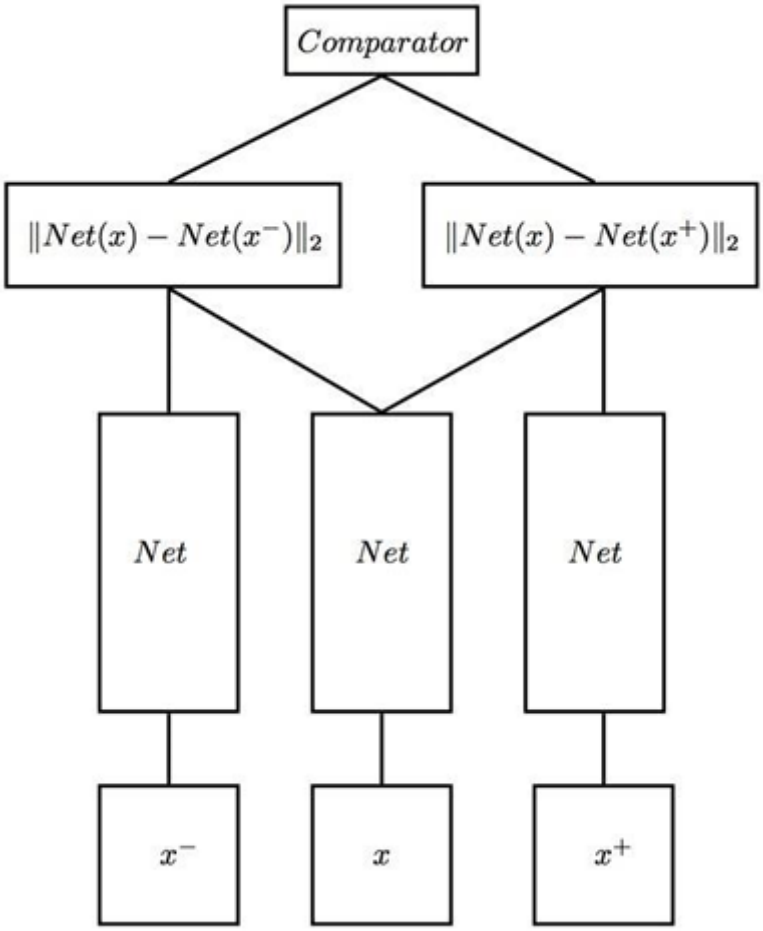
2. 三元网络（Triplet Network）

1）概述

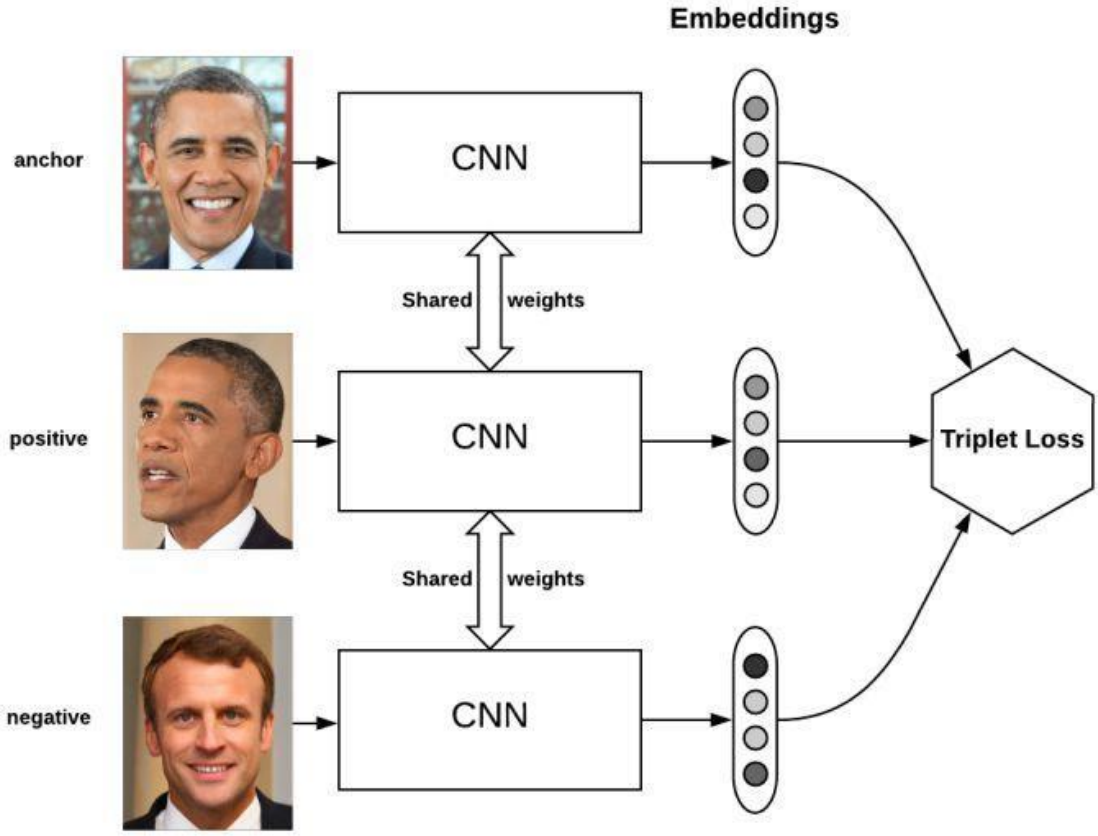
Triplet Network是Siamese Network的一种延伸，要解决的问题与Siamese Network的基本一致，主要用于单样本或样本较少的模型训练。与孪生网络不同的是，Triplet Network采用三个样本为一组：一个参考样本（也称为锚点样本），一个同类样本（正样本），一个异类样本（负样本）。Triplet Network网络将输入映射到某个特征空间中，使得参考样本与正样本距离足够小、与负样本足够大。

2）网络结构

Triplet Network结构图如下所示：



- Triplet Network由3个相同的前馈神经网络（彼此共享参数）组成；
- 输入x为参考样本，x^-为负面样本，x^+为正面样本；
- 网络会输出两个值：参考样本与负面样本、正面样本的特征向量的距离。



3）损失函数

Triplet Network的损失函数称为Triplet Loss（三元损失），包含参照样本、正面样本、负面样本之间的距离。$||f(A)-f(P)||^2$表示参照样本与正向样本的距离，$||f(A)-f(N)||^2$表示参照样本与负面样本的距离。我们希望参照样本与正向样本的距离足够小，与负面样本的距离足够大，即：

$$||f(A)-f(P)||^2 <= ||f(A)-f(N)||^2 \tag{1}$$

为了避免两个表达式为0的情况（同时为0也满足该表达式），需要作出一些调整，参照样本与正向样本的距离值比起与负面样本的距离需大于某个常数值，所以表达式变为：

$$||f(A)-f(P)||^2 + \alpha <= ||f(A)-f(N)||^2 \tag{2}$$

其中，α为一个超参数，称为间隔（margin）。将等式右边移项到左边：

$$||f(A)-f(P)||^2 - ||f(A)-f(N)||^2 + \alpha <= 0 \tag{3}$$

所以为了定义这个损失函数，我们取这个和0的最大值：

$$L(A,P,N) = max(||f(A)-f(P)||^2 - ||f(A)-f(N)||^2 + \alpha, 0) \tag{4}$$

在训练中，为了使得效果更好，通常选择更难训练的三元组(A,P,N)进行训练。所谓难训练的三元组就是选择的三元组$d(A,P)$很接近$d(A,N)$，即$d(A,P) \approx d(A,N)$，在训练时竭力使式子右边部分变大，左边部分变小，使得左右至少有一个α的距离。并且，比起随机选择的三元组，这样的三元组还可以增加模型的学习算法的计算效率。

五、人脸识别

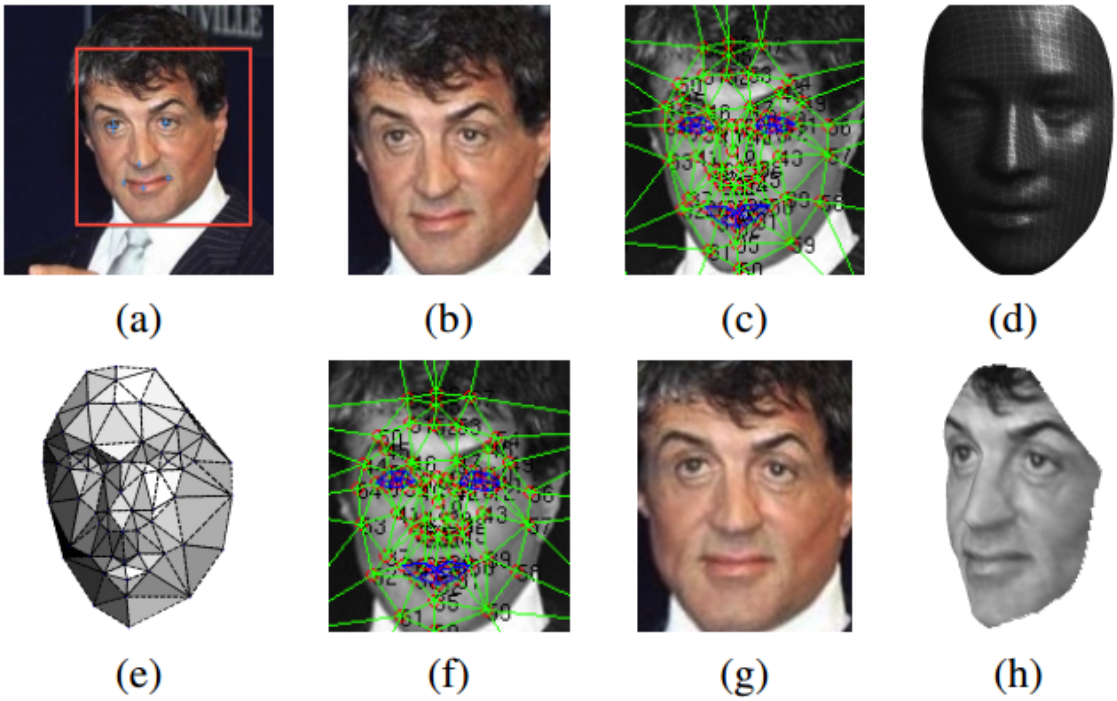
1. DeepFace（2014）

1）概述

DeepFace是Facebook研究人员推出的人脸验证模型，是深度学习技术应用于人脸识别的先驱。模型深度9层，超过1.2亿个参数。在LFW数据集上识别率达到97.25%，接近人类识别能力。

2）人脸对齐处理

和大多数模型一样，DeepFace采用基准点检测器指导对齐过程。在该模型中，使用了一个相对简单的基准点检测器，经过多次迭代来优化输出。在每一次迭代中，通过训练支持向量回归器（SVR）从图像描述中预测点的结构来提取基准点。

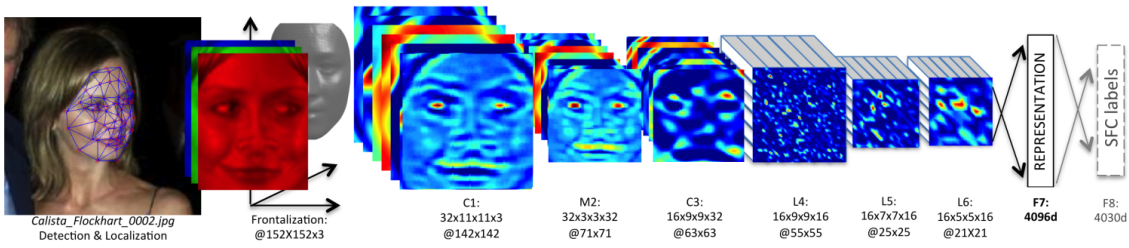


对齐步骤：

- 检测人脸6个基准点（眼睛2个、鼻尖1个、嘴巴3个）
- 裁剪人脸部分
- 在裁剪出的人脸中，使用67个基准点进行三角剖分
- 将二维对齐裁剪图像转换为三维参考形状
- 对三维形状进行旋转，生成正面二维图像

3）网络结构

DeepFace网络结构如下图所示：



- 输入：152*152经过预处理3D对齐的3通道面部图像
- 第一层：卷积层（论文中称为C1），采用32个11*11卷积核进行卷积，输出32个142*142的特征图

- 第二层：池化层（M2），以步幅为2执行最大池化操作，输出32个71*71的特征图
- 第三层：卷积层（C3），采用16个9*9卷积核进行卷积，输出16个63*63的特征图

以上三层主要提取底层特征，例如简单边沿、纹理。

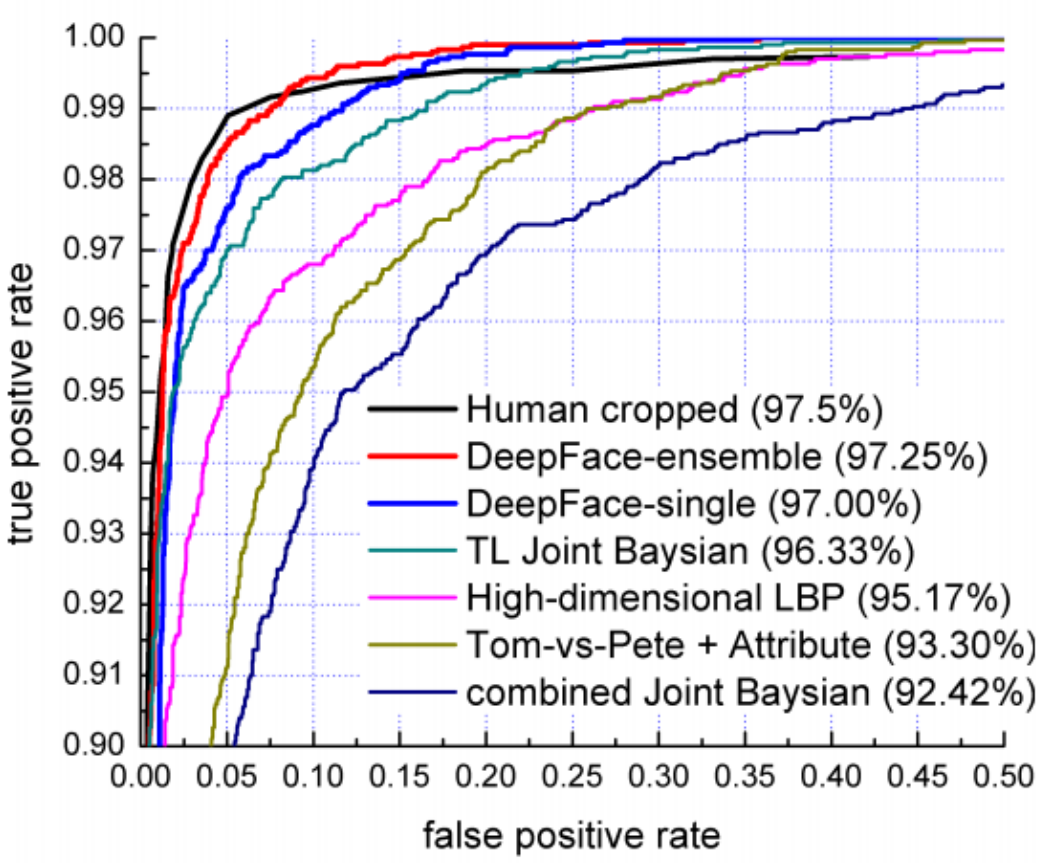
- 第四/五/六层：局部连接层（L4/L5/L6），滤波器组，像卷积层一样，应用滤波器组，但是特征映射中的每个位置都学习不同的滤波器组。局部层的使用不会影响特征提取的计算负担，但会影响训练参数的数量。
- 第七/八层：全连接层（F7/F8），用来捕获在面部图像的远处部分捕捉到的特征之间的相关性，例如眼睛的位置和形状以及嘴的位置和形状。最后一个全连接层在Softmax函数作用下产生K路输出。

4）效果

- **LFW数据集实验效果**

DeepFace在LFW数据集上实验准确率与效果图ROC（Receiver Operating Characteristic）曲线如下所示：

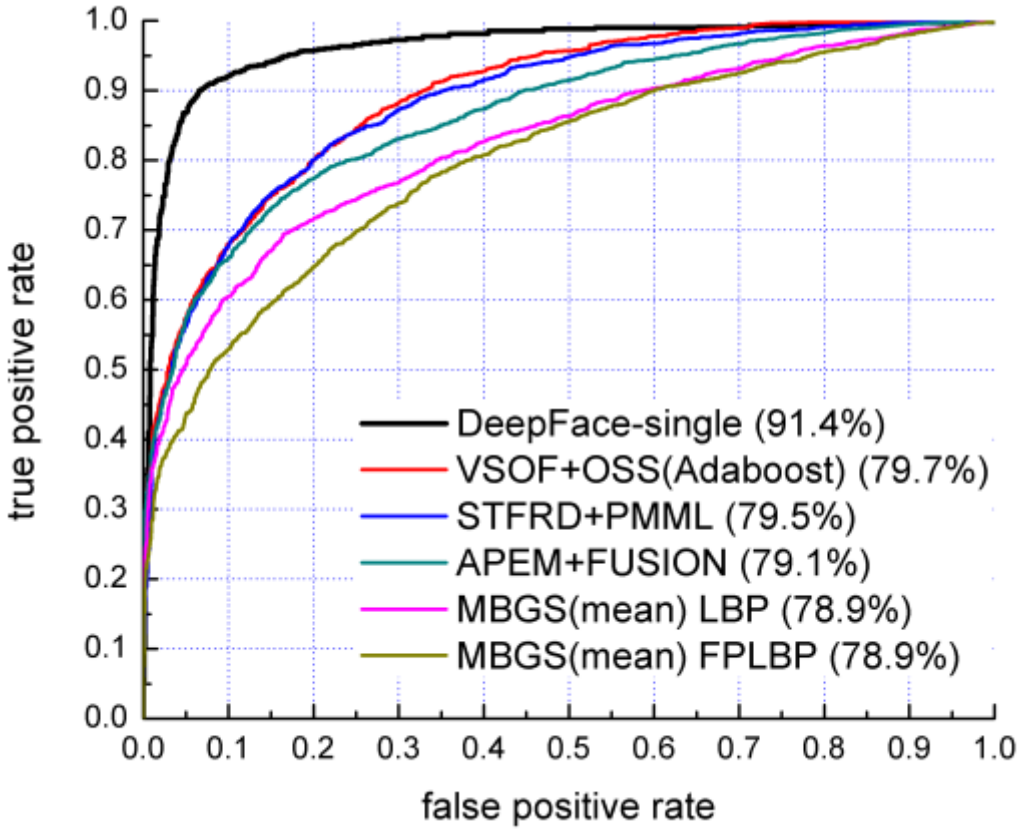
Method	Accuracy	Protocol
Joint Bayesian [6]	0.9242 ±0.0108	restricted
Tom-vs-Pete [4]	0.9330 ±0.0128	restricted
High-dim LBP [7]	0.9517 ±0.0113	restricted
TL Joint Bayesian [5]	0.9633 ±0.0108	restricted
DeepFace-single	**0.9592** ±0.0092	unsupervised
DeepFace-single	**0.9700** ±0.0087	restricted
DeepFace-ensemble	**0.9715** ±0.0084	restricted
DeepFace-ensemble	**0.9725** ±0.0081	unrestricted
Human, cropped	0.9753	



- **YTF数据集实验效果**

DeepFace还在视频级人脸验证数据集上进行了进一步验证。YouTube视频帧的图像质量通常比网络照片差，主要是由于运动模糊或观看距离。为每对训练视频创建50对帧，每个视频一对，并根据视频训练对标记这些帧是否相同。DeepFace取得了91.4%的准确度，将先前的最佳方法的误差减少了50%以上。如下图所示：

Method	Accuracy (%)	AUC	EER
MBGS+SVM- [30]	78.9 ±1.9	86.9	21.2
APEM+FUSION [21]	79.1 ±1.5	86.6	21.4
STFRD+PMML [9]	79.5 ±2.5	88.6	19.9
VSOF+OSS [22]	79.7 ±1.8	89.4	20.0
DeepFace-single	**91.4** ±1.1	96.3	8.6



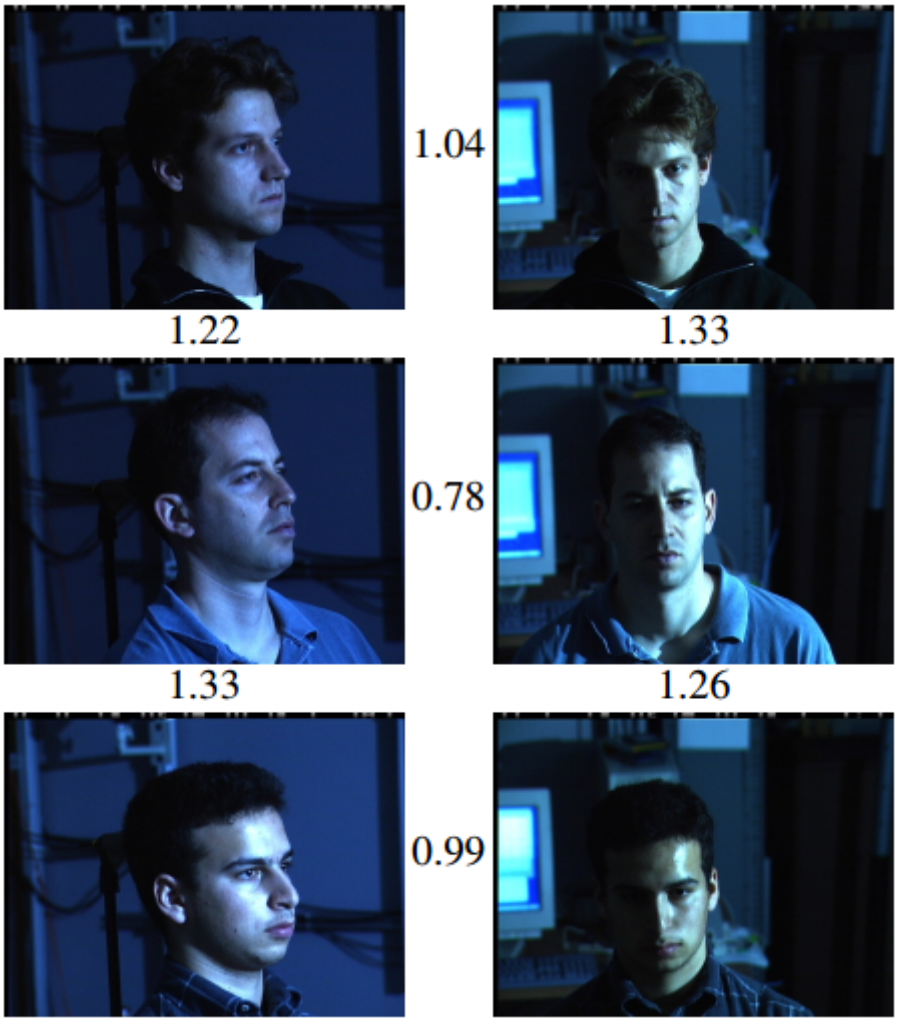
5）计算效率

单核Intel 2.2GHz CPU，DeepFace每幅图像运行0.33秒，包括图像解码、人脸检测和对齐、前馈网络和最终分类输出。

2. FaceNet（2015）

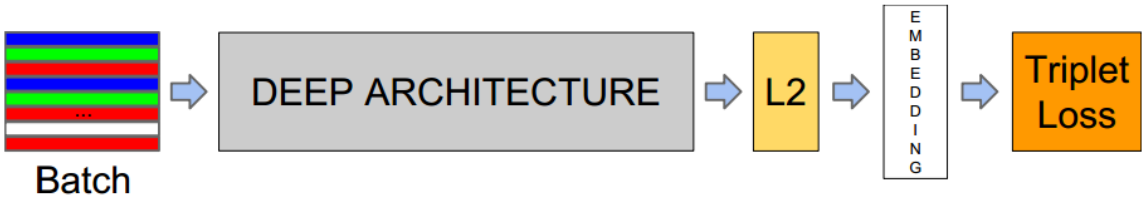
1）概述

FaceNet是Google研究人员2015年推出的人脸识比对模型。其思想是将人脸照片转换为128-D的向量，判断向量之间的距离（差异），如果距离大于某个值，则认为不是同一个人的人脸图像；如果距离小于某个值，则认为是同一个人的人脸图像。

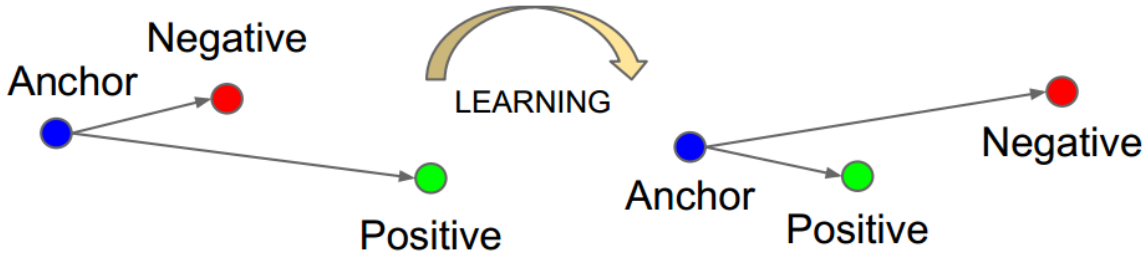


2）模型结构

FaceNet模型结构如下图所示：



- 批量输入层：输入一个批次的样本；
- 深度CNN：用来提取数据特征，然后进行L2标准化，输出向量特征（嵌入）。该部分可以采用不同的CNN模型；作者原论文中给出了两种CNN结构，并对性能进行了对比；
- 三元损失函数：接收特征向量，构建三元损失函数，使得Anchor-Positive之间的距离足够小，Anchor-Negtive之间的距离足够大。



3）损失函数

FaceNet采用三元损失函数，表达式如下：

$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha <= ||f(x_i^a) - f(x_i^n)||_2^2 \tag{5}$$

$$\forall f(x_i^a), f(x_i^p), f(x_i^n) \in \mathrm{T} \tag{6}$$

其中，x_i^a表示锚定样本（参考样本），x_i^p表示正样本，x_i^n表示负样本。
$||f(x_i^a) - f(x_i^p)||_2$表示参考样本与正样本间距离的L2范数（欧式距离）。T是训练集中所有可能的三元组的集合，具有基数N。训练优化的目标损失函数为：

$$\sum_i^N [\ ||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha\ ] \tag{7}$$

在训练过程中，为了让模型快速收敛，通常选择难训练的三元组样本进行训练，即$argmax\ x_i^p(||f(x_i^a) - f(x_i^p)||_2^2)$或$argmin\ x_i^n(||f(x_i^a) - f(x_i^n)||_2^2)$. 实际训练中，采用从小批量中选择argmin和argmax，而不是所有样本。

4）效果与性能

① 不同网络结构及参数

网络结构一：22层，140M（1.4亿）个参数，1.6B（16亿）次左右浮点运算

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	
total				140M	1.6B

网络结构二：基于GoogLeNet的初始模型。参数量减少了约20倍（约为6.6M~7.5M），每个图像需要220M次浮点运算。

② 不同图像质量的正确率

下表是不同压缩率的JPEG图像和不同分辨率图像的正确率：

jpeg q	val-rate
10	67.3%
20	81.4%
30	83.9%
50	85.5%
70	86.1%
90	86.5%

#pixels	val-rate
1,600	37.8%
6,400	79.5%
14,400	84.5%
25,600	85.7%
65,536	86.4%

③ 不同嵌入维度的正确率

#dims	VAL
64	$86.8\% \pm 1.7$
128	$87.9\% \pm 1.9$
256	$87.7\% \pm 1.9$
512	$85.6\% \pm 2.0$

④ 不同数据集上准确率

- LFW数据集：使用额外的人脸对齐平均准确率达99.63%
- YFB数据集：平均准确率95.18%