# 计算机视觉

## 概述

计算机视觉（Computer Vision, CV），顾名思义，就是让计算机能够像人类的眼睛那样，具有“看”和“分析”的功能，但是计算机不同于人类与生俱来的“看”和“分析”的能力，需要借助一定的科学技术来实现这些功能，这些技术包括数字图像处理、数字信号处理、光学、物理学、几何学、应用数学、模式识别、人工智能等等。计算机视觉技术的应用已经涉及到计算几何、计算机图形学、图像处理、机器人学等领域。在多媒体信息时代，对计算机视觉的研究工作显得尤为重要，人脸是图像中最重要的对象之一，人脸识别技术是计算机视觉与模式识别领域的重要课题。人脸识别技术是基于人的脸部特征信息进行身份识别的一种生物识别技术，这种技术自动在采集到的包含人脸的图像或者视频中检测并跟踪人脸，然后采用摄像机或摄像头采集含有人脸的图像或视频流，并自动在图像中检测和跟踪人脸对检测到的人脸进行脸部特征定位、提取，记忆存储和比对辨识，从而识别不同人的身份。在商业界，随着越来越多企业跨界进入人脸识别领域，势必加剧该领域的竞争。未来的人脸识别技术的应用领域会越来越广，市场也将更广阔。本文从人脸识别的角度重点阐述一下计算机视觉的技术框架、算法模型、开源项目以及发展瓶颈。

## 技术框架

### 概述

人脸识别（Face Recognition）技术是通过视频采集设备获取用户的面部图像，然后通过一系列的算法对用户的脸部特征进行提取和分析，并和事先处理好且进行过信息标注的人脸数据库进行对比，进而判断出用户的真实身份再利用核心的算法对其脸部的五官位置、脸型和角度进行计算分析，人脸识别技术基于局部特征区域的单训练样本人脸识别方法，大致分为三步（图1：）：第一步，人脸检测，就是对图像中对应于人脸的局部区域进行定义；第二步，人脸对齐，就是对处于运动状态的人的脸部进行姿态校正。第三步，人脸验证，就是通过把人脸的特征输入分类器，用投票或线性加权等方式得到终识别结果。



1. 人脸识别过程

### 人脸检测

人脸检测（Face Detection）是指对通过人脸扫描获得的任意的视频或图片，采用一定的策略对其进行搜索以确定其中是否含有人脸，并把与人脸相关的内容寻找出来，然后返回人脸的位置、大小和姿态。人脸检测的过程如下图所示：



1. 人脸检测过程

人脸检测的过程，简单来讲就是“扫描+判定”的过程，它的输入是一张图像，输出是人脸聚焦框，输出的聚焦框可以是零个，一个或者多个。比较形象的表述就是，在一张图像里面，把人脸用一个聚焦框框出来，这个聚焦框可以是正方形，矩形等等。

### 人脸对齐

因为人在不同的时间点上运动状态是不一样的，同一个人的人脸姿态会存在差异，在没有进行人脸识别算法之前，计算机会把这些存在差异的图片判断为不同的人。为了保证识别的准确率，就需要进行人脸对齐。人脸对齐就是将两个不同的形状进行归一化的过程，将一个形状尽可能地贴近另一个形状。具体就是将人脸进行旋转、缩放、抠取等一系列的姿态校正操作，将人脸调整到预定的大小和形态，然后自动定位出面部关键特征点，如眼睛、鼻尖、嘴角点、眉毛以及人脸各部件轮廓点等，以便进行后续的识别，提高识别的准确率。这些一系列的姿态校正操作也叫做仿射变换，人脸教程过程如下图所示：



1. 仿射变换

可以使用最小二乘法使得仿射变换后的所有点，和我们参照的目标点的距离和最小，使得这些点尽可能对齐，从而保证仿射变换的效果。定位面部关键点特征（图4：）是利用关键点附近的信息以及各个关键点之间的相互关系来定位，方法大致分为两类：一类是基于模型的方法，典型的方法是Cootes在1995年提出的ASM方法，这类方法使用的是PCA参数模型，在本章的1.3.2节我们会详细介绍这种模型；另一类是基于回归的方法，典型的是ESR (Explicit Shape Regression) 和ERT (Ensemble of Regression Trees) 算法。



1. 人脸对齐

### 人脸对比

人脸对比一般包括三个场景：1:1对比、1:N对比，M:N对比。

1:1对比就是对人的面部图像进行采集并将采集到的人脸图像与人像数据库进行人脸比对，并判别是否一致的过程，1:1对比可以简单理解为证明你就是你，IPhone的面容ID，登机、验票、支付都属于1:1对比，例如使用面容ID对苹果IPhone手机进行解锁的过程就是典型的1:1场景。1:1作为一种静态比对，在金融、信 息安全领域中有巨大的商用价值。

1:N 对比是在拥有大量数据的人像数据库中找出能够匹配待识别的人脸的数据。1:N能够进行动态对比，可以通过动态的视频流获取人脸数据，不需要待识别对象的主动配合，因此具有非强制性与高效性。公共安全管理与VIP客户人脸识别等都属于1:N对比场景，因为不需要人特意进行配合，采集的图像数据可能存在过度曝光，画面不清，侧脸，距离远等特点，动态1:N对比的难度要远高于静态的1:1对比。

M:N 对比是对采集到的图像中的所有人进行人脸识别并与人像数据库进行比对的过程。M:N也是一种动态对比，在公共安防，迎宾，机器人等领域有广泛应用。但是M:N模式也存在一定的挑战，因为一张图像中包含多个人脸信息，这样每个人脸的分辨率相对较低，再加上要同时识别的人数很多，所依赖的人脸数据库就及其庞大，因此M:N模式可能会产生很高的错误率从而降低可靠性。

以上三个方面就是人脸识别的主要技术框架，在不同的识别阶段采用的算法也不相同。在下一节我们重点看一下人脸识别基本的算法模型。

## 算法模型

### 概述

### 主成分分析

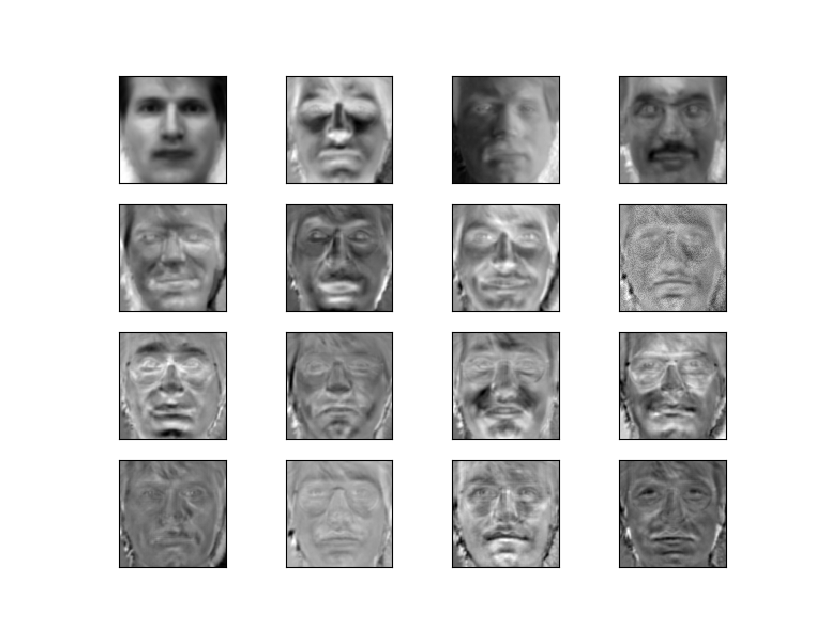
主成分分析（PCA, Principle Components Analysis）是模式识别领域一种重要的方法，它为我们提供了一种简单而有效的压缩图像的方法，PCA方法在将高维向量向低维向量转化时，使低维向量各分量的方差最大，且各分量互不相关，因此可以达到最优的特征抽取，已被广泛地应用于人脸识别算法中。

主成分分析用到了线性代数中的特征值分析，它能够通过数据的一般格式来解释数据的真实结构，也就是特征向量和特征值。PCA算法就是计算数据的协方差矩阵的特征值和特征向量，然后将特征值进行排序，并保留最大的N个特征值，最后将数据转换到这N个特征向量构建的新空间中。进行PCA处理后，数据特征从原来的维度降低到我们选择的N维空间中，从而降低了数据的复杂度和计算量，缺点是这个降维过程也可能导致有用信息的损失。

提取N个主成分的过程大致如下：

1. 去除平均值
2. 计算协方差矩阵
3. 计算协方差矩阵的特征值和特征向量
4. 将特征值从大到小排序
5. 保留最上面的N个特征向量
6. 将数据转换到保留的N个特征向量构建的新空间中。

图5：是使用Yale人脸库进行主成分分析的一个实例，这里从人脸库中选择15张人脸进行PCA分析，每张图片是由100x100个像素点组成的黑白图片，对每张这样的图片的数据逐行提取，可以得到一个10000维的向量。对这15张人脸图片数据进行平均，可算这些样本的均值向量 (10000维)，进行平均后的人脸图片如图5：中左上角第一个图片所示；然后根据样本均值向量计算出协方差矩阵 (10000x10000)。对这个协方差矩阵，可以得到其特征向量（10000x15），还原后的人的特征脸图片如图5：中的第二到第十六张图片所示。我们可以选取特征值最大的7个特征向量进行分析，计算出的前7个特征向量（对应最大的7个特征值）的特征脸图片如图6：（第一张是对15张人脸进行平均后的图片）。当数据的维数较大时，直接计算数据的特征值和特征向量很困难，利用人工神经网络的学习能力，可以通过训练，逐步进行主分量分析。训练后的网络权值作为数据矩阵的特征向量，网络的输出作为输入样本数据在低维空间各方向上的投影。在1.3.3节中我们将介绍一下人工神经网络算法。



1. 所有特征向量对应的特征脸



1. 前七个主成分对应的特征脸

### 人工神经网络算法

人工神经网络是一种非线性动力学系统，具有良好的自组织、自适应能力。目前神经网络方法在人脸识别中的研究方兴未艾。而神经网络方法可以通过学习的过程获得对人脸识别的许多规律或规则进行隐性的表达，它的适应性更强，一般也比较容易实现。人工神经网络识别速度快，但识别率低。

人工神经网络是由大量处理单元互联组成的非线性、自适应信息处理系统。它是在现代[神经科学](https://baike.baidu.com/item/%E7%A5%9E%E7%BB%8F%E7%A7%91%E5%AD%A6)研究成果的基础上提出的，试图通过模拟大脑神经网络处理、记忆信息的方式进行信息处理。人工神经网络的基本单元是神经元，它由生物神经网络抽象出来的一个模型（图7：，这个模型有三个基本要素：

1、一组连接（对应于生物神经元的突触），连接强度由各连接上的权值表示，权值为正表示激活，为负表示抑制。

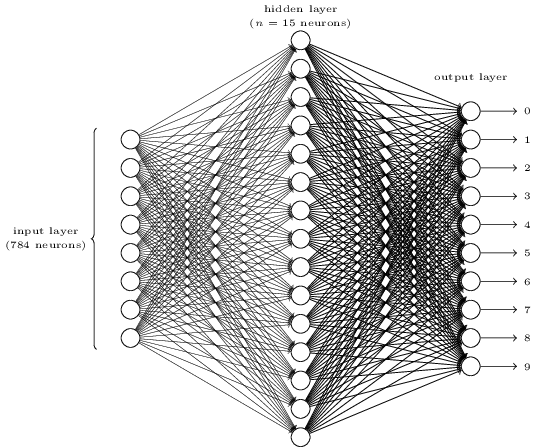
2、一个求和单元，用于求取各输入信号的加权和（线性组合）。

3、一个非线性激活函数，起非线性映射作用并将神经元输出幅度限制在一定范围内。



1. 基本神经元模型

常用的人工神经网络是由多层功能神经元组成的层级结构（如图8：每层神经元与下一层神经元全互连，神经元之间不存在同层链接，也不存在跨层连接。这样的神经网络结构通常称为“多层前馈神经网络”（multi-layer feedforward neural networks），其输入层神经元接收外界输入，隐层与输出层神经元对信号进行加工，最终结果由输出层神经元输出。这里输入层神经元仅接受输入，不进行函数处理，隐层与输出层包含功能神经元。神经网络学习的过程，就是根据训练数据来调整神经元之间的“连接权重”（connection weight）以及每个功能神经元的阈值，调整结束后的连接权重和阈值就是神经网络学习到的内容，可以用来对未进行过训练的新样本进行识别。图-7是用于mnist手写数据识别的单隐层人工前馈神经网络，这是一个三层的神经网络，每张手写字体图片有78x78个像素，因此输入层有784个神经元，隐藏层由15个神经元组成，输出层有10个神经元，用来对应0~9十个数字。



1. 单隐层前馈网络举例

人工神经网络的训练通常使用梯度下降（Gradient Descent）算法来计算出合适的连接权重和阈值，并使用BP（Backpropagation）算法来提高算法的精度。在人脸识别的应用中，因为数据量庞大，对精度的要求更高，人工神经网络往往使用层数更多，结构更复杂的的模型，这种模型称之为深度学习模型。在1.4.2节我们将介绍一下深度学习的框架。

## 开源项目

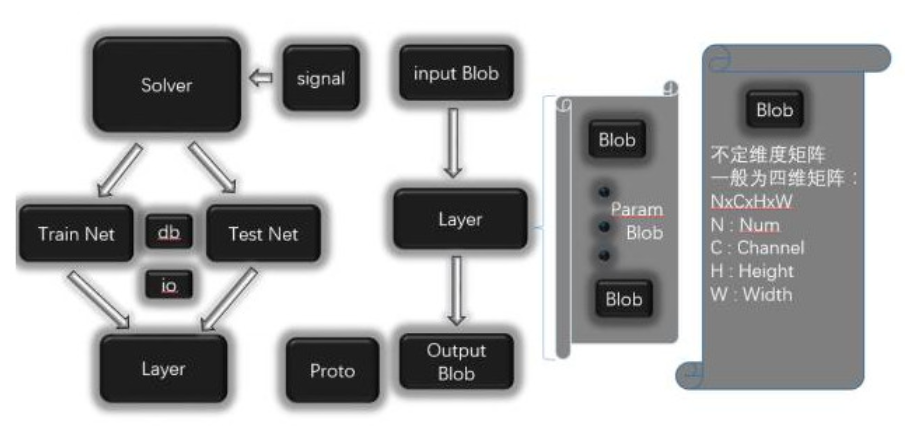
### 概述

人工智能时代，基于深度学习的计算机视觉领域有很多的开源的项目，这些项目按照其功能特点，可以划分为三类：一、基础技术平台和深度学习框架，如深度学习平台caffe；二、开源的计算机视觉相关图像资源库，如人脸识别资源库等；三、开源的计算机视觉算法模型，如人脸检测，人脸对齐和人脸识别算法模型等等。下面就从这三个方面进行深入的介绍。

### 深度学习框架

深度学习研究的热潮持续高涨，各种开源深度学习框架也层出不穷，其中包括TensorFlow、Caffe、Keras、CNTK、Torch7、MXNet、Leaf、Theano、DeepLearning4、Lasagne、Neon，等等。在计算机视觉领域最流行的的工具包是caffe（Convolutional Architecture for Fast Feature Embedding），caffe专注于图像处理，是第一个主流的工业级深度学习工具，由UC Berkely的Yangqing Jia编写和维护，采用卷积神经网络实现。Caffe是纯粹的C++/CUDA架构。支持命令行、Python和MATLAB接口；能够在CPU和GPU直接无缝切换。

Caffe框架主要有五个组件，Blob，Solver，Net，Layer，Proto，其结构图如图8：所示。Solver负责深度网络的训练，每个Solver中包含一个训练网络对象和一个测试网络对象。每个网络则由若干个Layer构成。每个Layer的输入和输出Feature map表示为Input Blob和Output Blob。Blob是Caffe实际存储数据的结构，是一个不定维的矩阵，在Caffe中一般用来表示一个拉直的四维矩阵，四个维度分别对应Batch Size（N），Feature Map的通道数（C），Feature Map高度(H)和宽度(W)。Proto则基于Google的Protobuf开源项目，是一种类似XML的数据交换格式，用户只需要按格式定义对象的数据成员，可以在多种语言中实现对象的序列化与反序列化，在Caffe中用于网络模型的结构定义、存储和读取。



1. Caffe源码总体架构图

### 图像资源库

深度学习需要大量的数据作为输入，人脸识别的数据可以直接使用现有的图像资源数据库。从而进行人脸识别模型的训练，测试与优化。常用的人脸识别库有Youtube Face，LFW(Labeling Faces Wild)，CelebFaces(A)，MegaFace，CASIA WebFace等等，具体描述和应用见表1：。这些数据库可以从其对应的官方网站上进行下载和使用，另外还有不少其他的人脸数据库，不在这里一一描述，读者可以根据自己的需要选择合适的数据库进行使用。

|  |  |  |
| --- | --- | --- |
| **数据库** | **描述** | **应用** |
| **YouTube Face** | 1,595个人 3,425段视频 | 非限制场景、视频 |
| **LFW** | 5k+人脸，超过10K张图片 | 标准的人脸识别数据集 |
| **CeleBrayA** | 200k张人脸图像40多种人脸属性 | 人脸属性识别 |
| **MegaFace** | 690k不同的人的1000k人脸图像 | 新的人脸识别评测集合 |
| **WebFace** | 10k+人，约500K张图片 | 非限制场景 |

1. 常用的人脸识别数据库

### 开源的算法模型

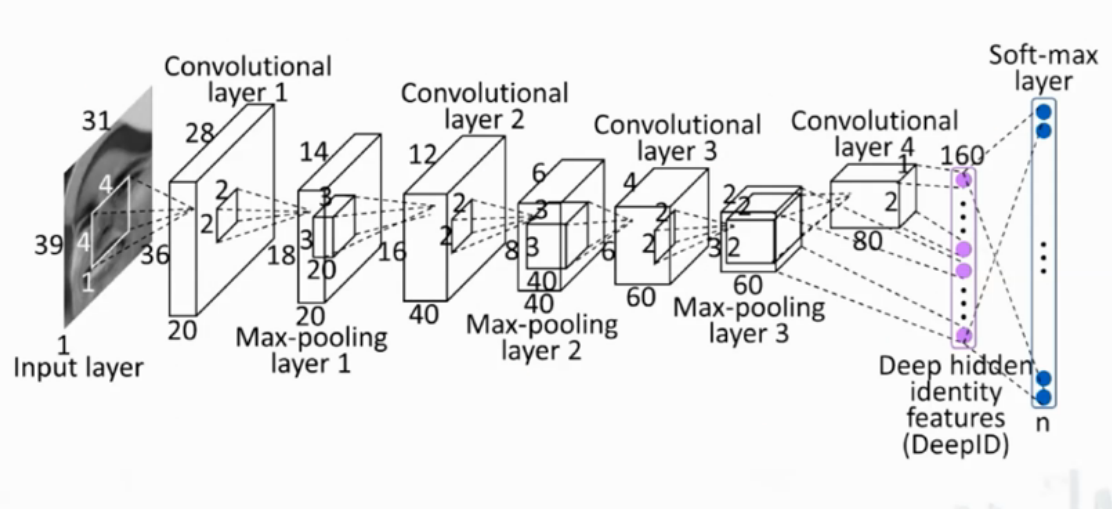
计算机视觉有很多开源的算法模型，这些算法模型基于特定的人脸识别样本会做相应的竞赛。表2：是ImageNet组织的人脸识别竞赛项目，表格中左边一列是算法模型，右边一列是使用该算法对应识别率。从表中可以看出，前期的准确率一般在84%-85%，到后期基本达到99%左右。图9：是一项人脸识别模型准确率的对比图，右边的小图是对左边识别率比较集中区域的一个放大。这些模型里面比较典型的是DeepID，旷视科技的face++，还有百度等等。一般情况下这些组织会公布参赛的算法模型的算法原理和逻辑，在开源社区里会有针对这些算法模型的实现，可以作为我们学习和参考的资源，并且可以复用到我们自己的场景里来。在实际操作中，我们可以增大训练样本的训练量，然后基于这个样本的训练量来训练自己的模型，再基于自己的模型适配到自己的真实场景里来。



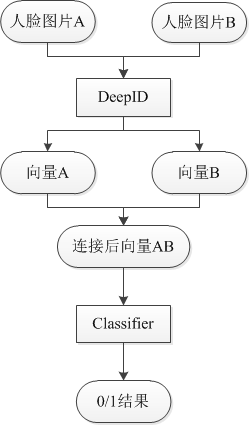
1. ImageNet人脸识别竞赛项目

下面我们对这些模型中的DeepID模型做一个详细的介绍:

DeepID是近几年认可度和效果取得比较好的一套机器视觉模型，它本身分为3个版本DeepID1、DeepID2和当前采用的DeepID3。DeepID一般用在人脸验证环节，人脸验证问题很容易就可以转成人脸识别问题，人脸识别就是多次人脸验证。DeepID采用的主要是卷积神经网络，通过将图像数据输入到神经网络学习到包含多维（使用LFW数据集的维数是160）不同特征的特征向量，然后用这些特征向量进行分类，一般包含输入层，卷积层和池化层。卷积层采用等价变换，池化层采用的是最大池化算法，最后输出适配结果。DeepID1的网络结构如图9：所示，这种网络结构和普通的卷积神经网络的结构相似，但是在倒数第二层，与Convolutional layer 4和Max-pooling layer3相连，由于卷积神经网络层数越高视野域越大，这样的连接方式可以既考虑局部的特征，又能考虑全局的特征。DeepID1的算法流程见图10：



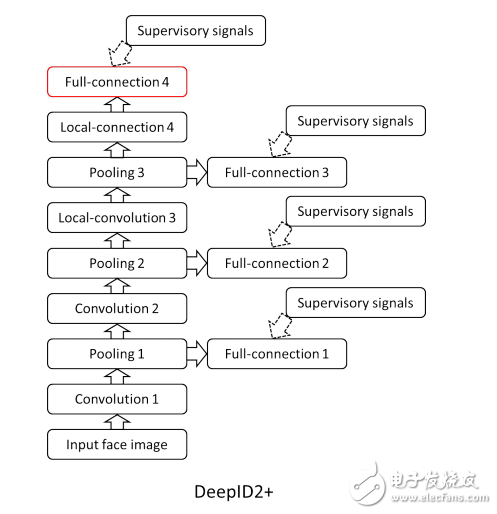
1. DeepID1网络结构



1. DeepID1算法流程

DeepID2在DeepID1的基础上添加了验证信号，其性能相对于DeepID1有了较大的提高。具体就是：DeepID1最后一层softmax使用的是Logistic Regression作为最终的目标函数，也就是识别信号，而DeepID2的最后一层目标函数上添加了验证信号，两个信号使用加权的方式进行了组合。

DeepID3继续更改了网络结构，DeepID的维数有了很大的提高（160维提高到512维），DeepID层不仅和第四层和第三层的max-pooling层连接，还连接了第一层和第二层的max-pooling层，并且对卷积神经网络进行了大量的分析，发现了几大特征，第一个特征是神经单元的适度稀疏性，该性质甚至可以保证即便经过二值化后，仍然可以达到较好的识别效果；第二个特征是高层的神经单元对人比较敏感，即对同一个人的头像来说，总有一些单元处于一直激活或者一直抑制的状态；第三个是 DeepID3的输出对遮挡非常鲁棒。DeepID3的网络结构如图11：所示。



1. DeepID3的网络结构

## 发展瓶颈

计算机视觉技术经过最近这些年的发展，取得了巨大的成就，技术水平有了很大的突破，在应用场景和识别精度上有了显著的提高。但是在取得巨大成就的同时，计算机技术的发展应用也存在一些瓶颈，一个例子是使用照片来进行人脸验证仍然能通过人脸验证系统，这就存在很大的安全隐患，这就产生了相应的应对措施，比如活体监测等等。但是实验证明，使用软件对人脸进行动态的视频合成仍然能攻破活体检测系统，这就需要更高级别的防御措施，从而推动这领域技术的发展。

## 参考书目

[1] Peter Harrrington，Machine Learning in Action

[2] 约书亚·本吉奥等著，赵申建等译，深度学习

[3] 周志华，机器学习