第九章 计算机视觉

计算机视觉是一门研究如何赋予机器“看”的智能的科学，更进一步的说，就是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等机器视觉，并进一步做图形处理。作为一个科学学科，计算机视觉研究相关的理论和技术，试图建立能够从图像或者多维数据中获取“信息”的人工智能系统。这里的信息指可以用来帮助做一个“决定”的信息。作为一个新兴学科，计算机视觉是通过对相关的理论和技术进行研究，从而试图建立从图像或多维数据中获取信息的人工智能系统。

**9.1 计算机视觉概述**

计算机视觉是从图像或者视频中提出符号或者数值信息，分析计算该信息以进行目标的识别、检测和跟踪等。更形象的说，计算机视觉就是让计算机像人类一样能看到图像，并看懂理解图像。

计算机视觉开始于20世纪50年代，主要用于分析和识别二维图像，如光学字符识别、显微图片的分析解释等。到60年代，通过计算机程序可以将二维图像转换成三维结构进行分析，从此开启三维场景下计算机视觉研究。到70年代，麻省理工学的人工智能实验室院首次开设计算机视觉课程，由著名的Horn教授主讲，同实验室的Marr教授首次提出表示形式（representation）是视觉研究最重要的问题。到80，90年代，计算机视觉迅速发展，形成感知特征的新理论框架并逐渐应用到工业环境中。到21世纪，计算机视觉领域呈现许多新的趋势，计算机视觉与计算机图形学深度结合，基于计算机视觉的应用也呈爆炸性增长，除了在手机、电脑上的应用，计算机视觉技术在交通、安防、医疗、机器人上有各种各样形态的应用。目前，计算机视觉技术已经应用在制造业、工业检验、文档分析、医疗诊断、军事目标跟踪、自主导航等系统当中。

现如今，随着数字化时代的到来，计算机视觉已广泛应用于各种各样的实际应用中，下面列出了八种不同的计算机视觉任务。

**图像分类**：图像分类根据图像的语义信息来区分不同类别的图像。它是计算机视觉中的一个重要的基础问题，也是目标检测、图像分割、目标跟踪、行为分析和人脸识别等高级视觉任务的基础。图像分类在许多领域都有着广泛的应用。例如：安防领域的人脸识别和智能视频分析等，交通领域的交通场景识别，互联网领域基于内容的图像检索和相册自动归类，医学领域的图像识别等。

**目标检测，跟踪和定位**：目标检测任务的目标是给定一张图像或一个视频帧，让计算机找出其中所有目标的位置，并给出每个目标的具体类别。对于人类来说，目标检测是一个非常简单的任务。然而，计算机能够“看到”的是图像被编码之后的数字，因此对于计算机而言很难解图像或视频帧中出现了人或是物体这样的高层语义概念，也就更加难以定位目标出现在图像中哪个区域。与此同时，由于目标会出现在图像或是视频帧中的任何位置，目标的形态千变万化，图像或是视频帧的背景千差万别，诸多因素都使得目标检测对计算机来说是一个具有挑战性的问题。

**图像语义分割**：图像语意分割，顾名思义是将图像像素按照表达的语义含义的不同进行分组/分割。图像语义是指对图像内容的理解，例如，能够描绘出什么物体在哪里做了什么事情等，分割是指对图片中的每个像素点进行标注，标注属于哪一类别。近年来用在无人驾驶技术中分割街景来避让行人和车辆、医疗影像分析中辅助诊断等。

**场景文字检测与识别：**许多场景图像中包含着丰富的文本信息，对理解图像信息有着重要作用，能够极大地帮助人们认知和理解场景图像的内容。场景文字识别是在图像背景复杂、分辨率低下、字体多样、分布随意等情况下，将图像信息转化为文字序列的过程，可认为是一种特别的翻译过程，即将图像输入翻译为自然语言输出。场景图像文字识别技术的发展也促进了一些新型应用的产生，如通过自动识别路牌中的文字帮助街景应用获取更加准确的地址信息等。

**图像生成：**图像生成是指根据输入向量，生成目标图像。这里的输入向量可以是随机的噪声或用户指定的条件向量。具体的应用场景有：手写体生成、人脸合成、风格迁移、图像修复、超分重建等。

**人体关键点检测：**人体关键点检测，通过人体关键节点的组合和追踪来识别人的运动和行为，对于描述人体姿态，预测人体行为至关重要，是诸多计算机视觉任务的基础，例如动作分类，异常行为检测，以及自动驾驶等等，也为游戏、视频等提供新的交互方式。

**视频分类：**视频分类是视频理解任务的基础，与图像分类不同的是，分类的对象不再是静止的图像，而是一个由多帧图像构成的、包含语音数据、包含运动信息等的视频对象，因此理解视频需要获得更多的上下文信息，不仅要理解每帧图像是什么、包含什么，还需要结合不同帧，知道上下文的关联信息。

**度量学习：**度量学习也称作距离度量学习、相似度学习，通过学习对象之间的距离，度量学习能够用于分析对象时间的关联、比较关系，在实际问题中应用较为广泛，可应用于辅助分类、聚类问题，也广泛用于图像检索、人脸识别等领域。

计算机视觉在许多领域有着广泛的应用价值。据说人一生中70%的信息是通过“看”来获得的，不难想到，看的能力对人工智能是非常重要的。大部分的AI系统都需要和人交互或者需要根据周边环境情况做决策,因此，计算机“看”的能力的重要性就更为凸显。随着计算机视觉的发展,越来越多的计算机视觉系统开始走入人们的日常生活，如人脸识别、指纹识别、车牌识别、视频监控追踪、自动驾驶汽车、增强现实等。

计算机视觉是让计算机理解图像的过程。图像处理能力赋予计算机看的能力以用于获取信息，图像是人工智能的重要输入。图像处理的目的是将低质量的输入图像转换为高质量的输出图像。常用的方法包括图像压缩编码、图像变换、图像描述、图像增强和图像复原。图像压缩编码是为了减少描述图像的比特数，以节省传输和存储的消耗。图像变换的目的是为了减少计算量，如将空间域的图像阵列变换到频域空间来处理。图像描述是图像理解的前提。它的功能是挖掘出描述图像的一般或主要信息。图像增强和图像复原主要用于改善图像质量，如去除噪声、增强高频信息等。上述图像处理技术主要依靠一些数学变换。

模式识别、机器学习、深度学习等算法赋予了计算机看懂的能力，是人工智能的关键与核心，更形象的说就是这些算法可以让计算机像人脑一样去理解图像。模式识别、机器学习、深度学习是让机器感知或学习的工具或方法。让计算机看懂的这一过程，就是根据图像或者视频数据建模的过程。建模则是用数学符号或者公式推理数据之中的一般模式或者规律，从而可以对新输入的数据进行分类或者回归，分类就是输出数据的类别，回归与数学中的映射函数类似，输出数据的可能值。

**9.2 数字图像的类型和表示**

由于计算机所“看到”的数字图像有一个个像素点组成，对于人而言看似很简单的计算机视觉任务对机器却极富挑战性。在计算机中，按照颜色和灰度的多少可以将图像分为二值图像、灰度图像、索引图像和真彩色RGB图像四种基本类型。目前，大多数图像处理软件都支持这四种类型的图像。

**二值图像：**即图像上的每一个像素只有两种可能的取值或灰度等级状态，0和1，0代表黑，1代表白，或者说0表示背景，而1表示前景。其保存也相对简单，每个像素只需要1Bit就可以完整存储信息。如果把每个像素看成随机变量，一共有N个像素，那么二值图有2的N次方种变化，而8位灰度图有255的N次方种变化，8位三通道RGB图像有255255255的N次方种变化。也就是说同样尺寸的图像，二值图保存的信息更少。二值图像通常用于文字、线条图的扫描识别（OCR）和掩膜图像的存储。例如图1（a）。

**灰度图像：**是每个像素只有一个采样颜色的图像，这类图像通常显示为从最暗的黑色到最亮的白色的灰度，尽管理论上这个采样可以任何颜色的不同深浅，甚至可以是不同亮度上的不同颜色。灰度图像与黑白图像不同，在计算机图像领域中黑白图像只有黑色与白色两种颜色；但是，灰度图像在黑色与白色之间还有许多级的颜色深度。灰度图像经常是在单个电磁波频谱如可见光内测量每个像素的亮度得到的，用于显示的灰度图像通常用每个采样像素8位的非线性尺度来保存，这样可以有256级灰度（如果用16位，则有65536级）。例如图1（b）。

**彩色图像：**指RGB真彩色图像，通常包含三个通道的信息，每个像素通常是由红（R）、绿（G）、蓝（B）三个分量来表示的，分量介于[0，255]。RGB图像每一个像素的颜色值（由RGB三原色表示）直接存放在图像矩阵中，由于每一像素的颜色需由R、G、B三个分量来表示，M、N分别表示图像的行列数，三个M×N的二维矩阵分别表示各个像素的R、G、B三个颜色分量。例如图1（c）。

**索引图像：**文件结构比较复杂，除了存放图像的二维矩阵外，还包括一个称之为颜色索引矩阵MAP的二维数组。MAP的大小由存放图像的矩阵元素值域决定，如矩阵元素值域为[0，255]，则MAP矩阵的大小为256×3，用MAP=[RGB]表示。MAP中每一行的三个元素分别指定该行对应颜色的红、绿、蓝单色值，MAP中每一行对应图像矩阵像素的一个灰度值，如某一像素的灰度值为64，则该像素就与MAP中的第64行建立了映射关系，该像素在屏幕上的实际颜色由第64行的[RGB]组合决定。索引图像一般用于存放色彩要求比较简单的图像，如Windows中色彩构成比较简单的壁纸多采用索引图像存放，如果图像的色彩比较复杂，就要用到RGB真彩色图像。



（a） (b) (c)

图1：（a)二值图像（b）256级灰度图（c）RGB真彩色图

除了上述四种图像，还有一类特殊的相机可以采集深度信息,即RGBD图像。RGBD图像对每个像素除了赋予红绿蓝彩色信息之外,还会有一个表达深度的值,即该像素距离摄像机的距离。其单位取决于相机的测量精度,一般为毫米，至少用2个字节表示。深度信息本质上反映了物体的3D形状信息。这类相机在体感游戏、自动驾驶、机器人导航等领域有潜在的广泛的应用价值。

此外,计算机视觉处理的图像或视频还可能来自人眼之外的成像设备，它们所采集的电磁波段信号超出了人眼所能够感知的可见光电磁波段范围,如红外、紫外、X光成像等。这些成像设备及其后续的视觉处理算法在医疗、军事、工业等领域有非常广泛的应用，可用于缺陷检测、目标检测、机器人导航等。例如在医疗领域，通过计算机断层X光扫描（CT)，可以获得人体器官内部组织的结构，3D的CT图中每个灰度值反映的是人体内某个位置对X射线的吸收情况，体现的是内部组织的致密程度。通讨CT图像处理和分析，可实现对病灶的自动检测和识别。

在大多数计算机视觉任务中，一般操作的都是RGB真彩色图像，或者为了研究方便，只针对一个通道进行研究，即灰度图像。

**9.3 常用计算机视觉模型和技术**

虽然计算机视觉的任务有多种，但是大多数任务本质上可以建模为一个广义的函数拟合问题，如图2所示。与任意的输入图像，需要学习一个函数,其中为一个需要学习的参数，使得，其中可能有两大类：

* 为类别标签，对应模式识别或机器学习中的“分类”问题，如场景分类、

图像分类、物体识别、精细物体类识别、人脸识别等视觉任务。这类任务的特点是输出为有限种类的离散型变量。

* 为连续变量或向量或矩阵，对应模式识别或机器学习中的“回归”问题,

如距离估计、目标检测、语义分割等视觉任务。在这些任务中，或者是连续的变量（如距离、年龄、角度等)，或者是一个回重（物体的横纵坐标位置和长宽)，或者是每个像素有一个所属物体类别的编号（如分割结果)。

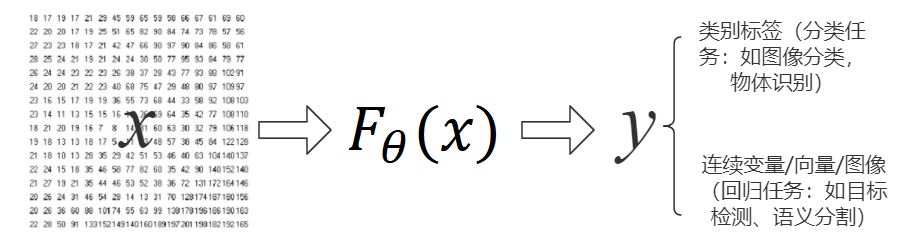


图2：常见视觉任务的实现方法模型

迄今为止，出现了很多方法来实现上述模型，其中大多数方法可以分为两大类，基于浅层模型和基于深度模型的方法。

**9.3.1 基于浅层模型的方法**

由于实现上述视觉任务的函数通常都是十分复杂的，因此，基于浅层模型的方法遵循“分而治之”的思想，将其拆分成多个子任务，分布求解。图3展示了一个典型的浅层视觉模型的实现流程。



图3：常用的浅层视觉模型处理流程

**步骤1：图像预处理过程p**。输入为图像*x*，输出为处理后的图像*x*’。主要操作包括目标对齐，几何归一化，亮度或颜色校正等。该过程的目的是提高数据的一致性。

**步骤2：特征设计与提取过程q**。从预处理后的图像中提取特征，这些特征一般是反映图像的边缘、部件或场景特性。此过程一般由人工根据专家知识进行设计。

**步骤3：特征汇聚或变换h**。为了方便后续分类或回归过程，需要对前一过程提取的局部特征*z*进行统计汇聚或降维处理，以得到维度更低的特征*z’*。例如线性模型，即*z’=Wz*，其中*W*为矩阵形式表达的线性变换，一般需要在训练集合进行学习得到。

**步骤4：分类器或回归器g**。采用机器学习或模式识别的方法，基于一个有类别标签的训练集通过监督学习的机器学习方法来学习得到。例如采用线性模型，即y= *Wz’*，可以通过优化得到，其中表示步骤3得到的*xi*的特征。

上述流程可以理解为将拆分成了四个序贯执行的四个函数*p, q, h, g*(即*y=g*(*h*(*q*(*p(x*))))。很显然，与深度学习端到端的自动学习相比，上述流程非常依赖基于专家知识的人工设计。通常称这些模型为浅层视觉模型。下面将针对后面三个步骤进行概述。

1. **特征设计与提取方法**

人工设计特征本质是一种专家知识驱动的方法，即研究者自己或通过咨询特定领域的专家，根据理解来设计流程提取自认为或专家认为“好”的特征。例如，在人脸识别研究初期，大部分研究人员都认为利用面部关键特征点的相对距离、角度或器官面积等就能实现人脸识别，但这些特征很快就被实验证明效果并不好。目前，大多数人工设计的特征有两大类，建模图像中全部像素或多个不同区域像素中所蕴含的信息的全局特征和只从一个局部区域的少量像素提取信息的局部特征。

全局特征通常对图片的颜色、整体结构或形状等进行建模，或者是对图像场景的空间形状属性进行建模。相比于局部特征，全局特征较为粗粒度，适用于高效而无须精分的任务场景。而局部特征可以用来提取更为精细的特征，因此具有普适性。在这些年的发展中，研究人员设计了大量的局部特征。这些局部特征通常以建模边缘、梯度或纹理等为目标，采用的方法包括滤波器设计、局部统计量计算、直方图等。目前最为典型的局部特征有SIFT、SURF、HOG、LBP、Gabor滤波器、DAISY、BRIEF、ORB、BRISK等数十种，下面以SIFT为例详细介绍其提取方法。

SIFT即尺度不变特征变换，是用于图像处理领域的一种描述。这种描述具有尺度不变性，可在图像中检测出关键点，是一种局部特征描述子。SIFT算法实现特征匹配主要有以下三大工序：

**步骤1：尺度空间的极值检测**。关键点是一些十分突出的不会因光照、尺度、旋转等因素而消失的点，比如角点、边缘点、暗区域的亮点以及亮区域的暗点。此步骤是搜索所有尺度空间上的图像位置。通过高斯微分函数来识别潜在的具有尺度和旋转不变的兴趣点。

尺度空间理论最早出现于计算机视觉领域，当时其目的是模拟图像数据的多尺度特征。尺度空间理论的主要思想是利用高斯核对原始图像进行尺度变换，获得图像多尺度下的尺度空间表示序列，对这些序列进行尺度空间特征提取。二维高斯函数定义如下:

（1）

一幅二维图像，在不同尺度下的尺度空间表示可由图像与高斯核卷积得到:

（2）

其中(*x, y*)为图像点的像素坐标，*I*(*x, y*)为图像数据,*L*代表了图像的尺度空间。称为尺度空间因子，它也是高斯正态分布的方差，其反映了图像被平滑的程度，其值越小表征图像被平滑程度越小，相应尺度越小。大尺度对应于图像的概貌特征，小尺度对应于图像的细节特征。因此，选择合适的尺度因子平滑是建立尺度空间的关键。

在这一步里面，主要是建立高斯金字塔和DOG(Difference of Gaussian)金字塔，然后在DOG金字塔里面进行极值检测，以初步确定特征点的位置和所在尺度。

(1)建立高斯金字塔

为了得到在不同尺度空间下的稳定特征点，将图像*I*(*x, y*)与不同尺度因子下的高斯核G(x, y, )进行卷积操作，构成高斯金字塔。

高斯金字塔有*o*阶，一般选择4阶，每一阶有*s*层尺度图像，*s*一般选择5层。在高斯金字塔的构成中要注意，第1阶的第1层是放大2倍的原始图像，其目的是为了得到更多的特征点;在同一阶中相邻两层的尺度因子比例系数是*k*,则第1阶第2层的尺度因子是*k*，然后其它层以此类推则可;第2阶的第l层由第一阶的中间层尺度图像进行子抽样获得，其尺度因子是*k2o*，然后第2阶的第2层的尺度因子是第1层的*k*倍即*k3*。第3阶的第1层由第2阶的中间层尺度图像进行子抽样获得。其它阶的构成以此类推。

(2)建立DOG金字塔

DOG即相邻两尺度空间函数之差，用D(x, y, )来表示，如下所示:

（3）

DOG金字塔通过高斯金字塔中相邻尺度空间函数相减即可，如图4所示。在图中，DOG金字塔的第1层的尺度因子与高斯金字塔的第1层是一致的，其它阶也一样。

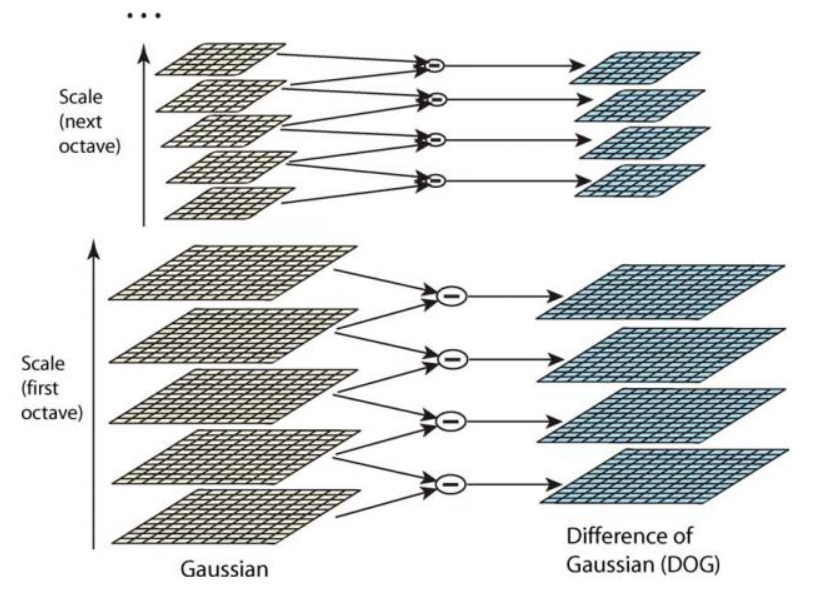


图4：高斯图像金字塔（S=2）与DOG金字塔

(3)DOG空间的极值检测

在上面建立的DOG尺度空间金字塔中,为了检测到DOG空间的最大值和最小值，DOG尺度空间中中间层(最底层和最顶层除外)的每个像素点需要跟同一层的相邻8个像素点以及它上一层和下一层的9个相邻像素点总共26个相邻像素点进行比较，以确保在尺度空间和二维图像空间都检测到局部极值，如图5所示。在图中，标记为叉号的像素如果比相邻26个像素的DOG值都大或都小，则该点将作为一个局部极值点，记下它的位置和对应尺度。

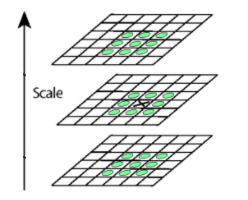


图5：DOG空间局部极值检测

**步骤2：特征点定位和特征方向赋值**。在每个候选的位置上，通过一个拟合精细的模型来确定位置和尺度。关键点的选择依据于它们的稳定程度。然后基于图像局部的梯度方向，分配给每个关键点位置一个或多个方向。所有后面的对图像数据的操作都相对于关键点的方向、尺度和位置进行变换，从而提供对于这些变换的不变性。

由于DOG值对噪声和边缘较敏感，因此，在上面DOG尺度空间中检测到局部极值点还要经过进一步的检验才能精确定位为特征点。下面对局部极值点进行三维二次函数拟和以精确确定特征点的位置和尺度，尺度空间函数*D*(*x*, *y*, σ)在局部极值点(*x*0,*y*0, σ)处的泰勒展开式如下所示。

（4）

其中*X*=(*x*, *y*, σ)T，一阶和二阶导数是通过附近区域的差分来近似求出的。通过对上述公式求导，并令其为0，得出精确的极值位置*X*max，如下所示：

（5）

在上面精确确定的特征点中，同时要去除低对比度的特征点和不稳定的边缘响应点，以增强匹配稳定性、提高抗噪声能力,这里不作详细介绍。

利用特征点邻域像素的梯度方向分布特性为每个特征点指定方向参数，使算子具备旋转不变性。

(6)

公式(6)为(*x, y*)处的梯度值和方向。*L*为所用的尺度为每个特征点各自所在的尺度，(*x, y*)要确定是哪一阶的哪一层。在实际计算过程中，在以特征点为中心的邻域窗口内采样，并用梯度方向直方图统计邻域像素的梯度方向。梯度直方图的范围是0°～360°，其中每10°一个柱，总共36个柱。梯度方向直方图的峰值则代表了该特征点处邻域梯度的主方向，即作为该特征点的方向。在梯度方向直方图中，当存在另一个相当于主峰值80%能量的峰值时，则将这个方向认为是该特征点的辅方向。一个特征点可能会被指定具有多个方向(一个主方向，一个以上辅方向)，这可以增强匹配的鲁棒性。

通过上面的两步,图像的特征点已检测完毕，每个特征点有3个信息:位置、对应尺度、方向。然后生成各个关键点的特征向量。

**步骤3：特征点匹配**。通过各关键点的特征向量，进行两两比较找出相互匹配的若干对特征点，建立景物间的对应关系。

首先，进行相似性度量。一般采用各种距离函数作为特征的相似性度量，如欧氏距离、马氏距离等。通过相似性度量得到图像间的潜在匹配。本文中采用欧氏距离作为两幅图像问的相似性度量。获取SIFT特征向量后，采用优先k-d树进行优先搜索来查找每个特征点的2近似最近邻特征点。在这两个特征点中，如果最近的距离除以次近的距离少于某个比例阈值，则接受这一对匹配点。降低这个比例阈值，SIFT匹配点数目会减少，但更加稳定。其次,消除错配。通过相似性度量得到潜在匹配对，其中不可避免会产生一些错误匹配，因此需要根据几何限制和其它附加约束消除错误匹配，提高鲁棒性。常用的去外点方法是RANSAC随机抽样一致性算法，常用的几何约束是极线约束关系。

1. **特征汇聚与特征变换方法**

通过步骤2提取的特征通常非常多，加剧了后续计算的难度。因此，在将特征输入给分类器或回归器之前需要对其进行进一步处理，将高维度特征降维或是变换到具有更好判别能力的新空间，即特征汇聚与特征变换。

特征汇聚方法最为典型的包括视觉词袋模型、Fisher可量和局部聚合向量（VLAD）方法。其中,词袋模型（BOW）最早出现在自然语言处理（NLP)和信息检索（IR）领域。该模型忽略掉文本的语法和语序,用一组无序的单词来表达一段文字或一个文档。受此启发,研究人员将词袋模型扩展到计算机视觉中,并称之为视觉词袋模型(BOVW)。简而言之，图像可以被看作文档,而图像中的局部视觉特征可以看作是单词的实例，从而可以直接应用BOW方法实现大规模图像检索等任务。

特征变换方法，又称子空间分析法。这类方法特别多，典型的方法包括主成分分析（PCA)、线性判别分析、核方法、流形学习等。其中,主成分分析是一种在最小均方误差意义下最优的线性变换降维方法，在计算机视觉中应用极为广泛。PCA在寻求降维变换时的目标函数是重构误差最小化，与样本所属类别无关,因而是一种无监督的降维方法。但在众多计算机视觉应用中,分类才是最重要的目标，因此以最大化类别可分性为优化目标寻求特征变换成为一种最自然的选择，这其中最著名的就是费舍尔线性判别分析方法FLDA。FLDA也是一种非常简单而优美的线性变换方法,其基本思想是寻求一个线性变换,使得变换后的空间中同一类别的样本散度尽可能小,而不同类别样本的散度尽可能大,即所谓“类内散度小,类间散度大”。核方法曾经是实现非线性变换的重要手段之一。核函数实现了一种隐式的非线性映射，将原始特征映射到新的高维空间，从而可以在无须显式得到映射函数和目标空间的情况下，计算该空间内模式向量的距离或相似度,完成模式分类或回归任务。前述的PCA和FLDA均可以Kernel化，以实现“非线性”的特征提取。

1. **分类器或回归器**

通过前两个步骤提取特征并对其进行进一步汇聚和变换，剩下的步骤就是分类器或回归函数的设计和学习了。计算机视觉中的分类器基本都借鉴模式识别或机器学习领域，例如最近邻分类器、线性感知机、决策树、随机森林、支持向量机、AdaBoost、神经网络等都可行。

**9.3.1 基于深度模型的方法**

计算机视觉应用深度学习的热潮爆发在2012年ImageNet比赛，Hinton教授研究组设计了深度卷积神经网络（DCNN）模型AlexNet相比于传统方法有飞跃性的提升。事实上，深度学习中的深度卷积神经网络也是通过滤波器提取局部特征，然后通过逐层卷积和汇聚,逐渐将“小局部”特征扩大为“越来越大的局部”特征,甚至最终通过全连接形成“全局特征”。但与浅层模型相比,深度模型的滤波器参数（权重）不是人为设定的,而是通过神经网络的BP算法等训练学习而来;而且DCNN模型以统一的卷积作为手段，实现了从小局部到大局部（即所谓层级感受野）特征的提取。与浅层模型相比,深度模型的滤波器参数（权重）不是人为设定的,而是通过神经网络的BP算法等训练学习而来的，而且DCNN模型以统一的卷积作为手段，实现了从小局部到大局部（即所谓层级感受野）特征的提取。

1. **基于深度模型的目标检测技术**

目标检测是计算机视觉中的一个基础问题，通俗地讲，就是给定一个输入的图像，我们希望模型可以分析这个图像里究竟有哪些物体，并能够定位这些物体在整个图像中的位置，对于图像中的每一个像素，能够分析其属于哪一个物体。

R-CNN是最早成功应用到目标检测上的深度模型。首先，R-CNN的输入是一个图片，输出是一个“选定框”和对应的标签。R-CNN采用了一种直观的方法来生成选定框：尽可能多地生成选定框，然后来看究竟哪一个选定框对应了一个物体。具体来说，针对图像，R-CNN先用不同大小的选定框来扫描，并且尝试把临近的具有相似色块、类型、密度的像素都划归到一起去。然后，再利用一个AlexNet的变形来对这些待定的选定框进行特征提取。在模型的最后一层，R-CNN加入了一个支持向量机（SVM）来判断待选定框是否是某个物体。判断好了选定框以后，R-CNN再运行一个线性回归来对选定框的坐标进行微调。

R-CNN虽然证明了在物体识别这样的任务中，CNN的确可以超越传统的模型，但整个模型由多个模块组成，相对比较繁琐。Fast R-CNN的一个重要特点就是观察到每一个待定的选定框都需要进行特征提取。这里的特征提取其实就是一个神经网络，往往非常消耗资源。而且很多待定的选定框有很多重叠的部分，可以想象就会有很多神经网络的计算是重复多余的。Fast R-CNN的另外一个特点就是尝试用一个神经网络架构去替代R-CNN中间的四个模块。Fast R-CNN和R-CNN相比在效果上差不多，但是训练时间快了9倍以上。

1. **基于全卷积网络的图像分割**

对于像素级的分类和回归任务（如图像分割或边缘检测)，代表性的深度网络模型是全卷积网络(fully convolutional network，FCN)。经典的DCNN在卷积层之后使用了全连接层，而全连接层中单个神经元的感受野是整张输入图像，破坏了神经元之间的空间关系，因此不适用于像素级的视觉处理任务。为此，FCN去掉了全连接层，代之以1×1的卷积核和反卷积层，从而能够在保持神经元空间关系的前提下，通过反卷积操作获得与输入图像大小相同的输出。进一步,FCN通过不同层、多尺度卷积特征图的融合为像系级的分类和回归任务提供了一个高效的框架。

1. **基于深度模型的视觉问答**

视觉问答（VQA）即使系统能够根据图片回答一些特定的问题。通常来说，一些具备完成人类任务的系统似乎总被认为是在科幻小说中才能出现的，而不会马上想到人工智能。然而，随着深度学习的出现，我们在视觉问答方面已经取得了巨大的研究进展。视觉问答中的算法大致可分为三个步骤：从图像中提取特征；从问题中提取特征；结合图像和文本特征来生成答案。在结合图像特征的情况下，在ImageNet上预先训练的卷积神经网络是最常见的选择。对于问题的生成，方法通常将问题建模为分类任务。目前，许多用于VQA的模型往往只使用了CNN的特征，CNN的特征大部分情况下是直接用的ImageNet训练好的模型，但由于用户的问题是不确定的，要正确回答他提出的任意的问题，需要解决多个计算机的任务，使用的特征有时过于单一。

**9.4 本章小结**

从20世纪50年代以来，计算机视觉取得了长足的进步，尤其是2012年以来,随着深度学习的发展，配合强监督大数据和高性能计算装置，众多计算机视觉算法的性能出现了质的飞跃。特别是在图像分类、人脸识别、目标检测、医疗读图等任务上逼近甚至超越了普通人类的视觉能力。

计算机视觉的多数任务可以归结为作用于输人图像的映射函数拟合期望输出的分类和回归问题。浅层视觉模型采用分而治之的思想，将映射函数拆分成预处理、特征提取、特征变换、分类和回归等步骤。但此类模型局限于人工经验设计或普遍采用简单的线性模型，在实际应用中不能得以广泛使用。以深度卷积神经网络为代表的深度学习模型克服了上述困难，采用层级卷积、逐级抽象的多层神经网络，实现了从输入图像到期望输出的、高度复杂的非线性函数映射。该模型大大提高了处理视觉任务的精度,而且显著降低了人工经验在算法设计中的作用,更多依赖于大量数据,让数据自己决定最“好”的特征或映射函数是什么。可以说，实现了从“经验知识驱动的方法论”到“数据驱动的方法论”的变迁。

计算机视觉领域自问世以来已经蓬勃发展，但直到机器能够在没有任何人类干预的情况下做出自己的决定之前，这一领域还有待发展。我们进一步的目标是探索新的技术和方法，降低计算成本，提高模型效果。