计算机视觉

1. 引言

计算机视觉，顾名思义，分为计算机和视觉两部分，视觉对我们人类来说是非常重要，不可替代的，我们可以借助视觉识别物体，辨别其所处的空间位置、远近等。诚然，这对于人类来说似乎是与生俱来的能力，然而如何让一个机器人、一台计算机拥有同样的“视觉”却是一件困难的事。计算机视觉的两个核心问题是重建与识别，即根据一组图像构建世界模型以及根据视觉和其他信息在遇到的对象之间进行区分，其基于模型的视觉方法主要有两种：对象模型与渲染模型。本文将从图像形成、基本图像属性、图像分类、目标检测、3D 世界、计算机视觉应用 6 个方面展开介绍。

2. 图像形成

成像会扭曲物体的外观，带来缩短、生长等效应，这些效应的模型对于建立有能力的物体识别系统至关重要，也为重建几何体提供了强有力的线索。主要的成像有以下三种：

2.1 针孔成像

其基本原理是到达传感器的每个光子都会产生电效应，其强度取决于光子的波长，输出是某个时间窗口内所有这些影响的总和，这意味着图像传感器报告到达光强度的加权平均值。该均值与波长、光子到达的方向、时间和传感器的面积有关。

为了看到聚焦的图像，我们必须确保到达传感器的所有光子都来自物体上大致相同的点，最简单额方法是使用针孔摄像头查看静止物体，当然我们也可以用针孔相机拍摄运动物体的聚焦图像，只要物体在传感器的时间窗口内只移动一小段距离。否则，运动物体的图像会散焦，这种效果称为运动模糊。

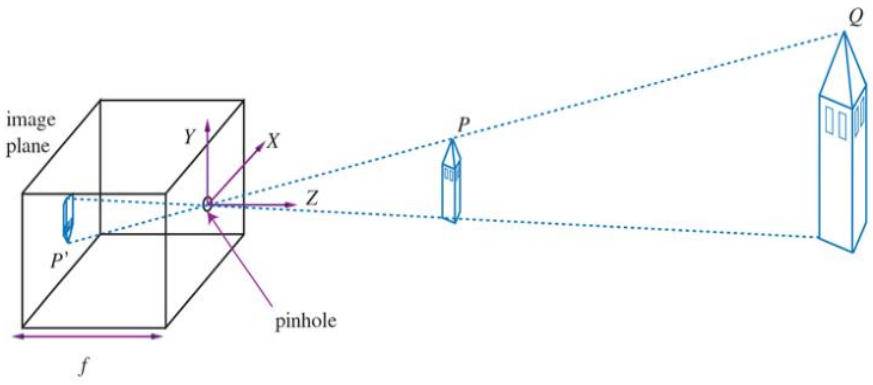


图 2-1 针孔成像

2.2 透视成像

针孔相机可以很好地聚焦光线，但因为针孔很小，只有一点光线会进入，图像会变暗。然而如果扩大孔（光圈）使图像更亮，照射到图像平面上某个特定点的光线将来自真实场景中的多个点，因此图像将散焦。脊椎动物的眼睛和现代相机使用镜头系统——眼睛中的一块透明组织或相机中的多个玻璃镜头元件系统来聚焦图像。在图 2-2 中，我们看到蜡烛顶端的光线向各个方向传播。相机（或眼睛）捕捉所有照射到镜头上任何地方的光都比针孔大得多，并将所有光聚焦到图像平面上的一个点上。来自蜡烛其他部分的光也会被收集起来，并聚焦到图像平面上的其他点，结果是一个更明亮、噪音更小、聚焦的图像。

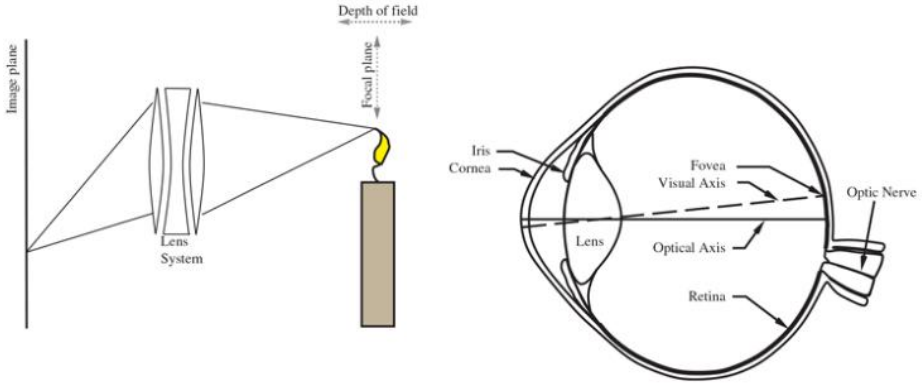


图 2-2 透镜成像

2.3 比例正交投影

透视成像的几何效果并不总是明显的，例如，街对面建筑物上的窗户看起来比附近的窗户小得多，但相邻的两扇窗户大小大致相同，即使其中一扇稍远。此时我们可以选择缩放正交投影的简化模型来处理窗口，如果对象上所有点的深度都在范围内，则透视比例因子可以近似为常数。缩放正交投影模型中仍会出现前缩，因为它是由对象偏离视图而引起的。

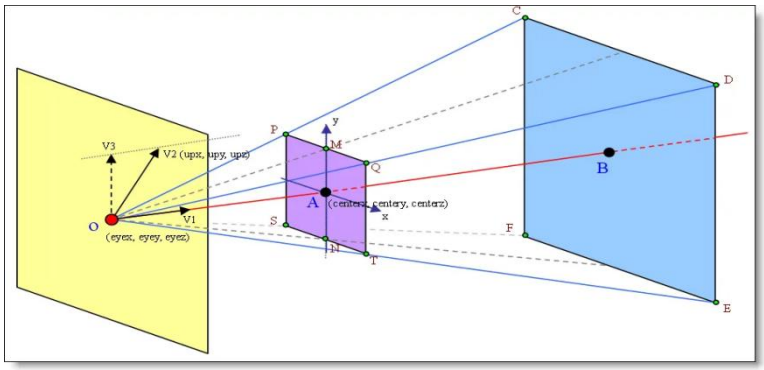


图 2-3 比例正交投影

2.4 光影与颜色

图像中像素的亮度是场景中投影到该像素的曲面片亮度的函数，其与环境光的整体强度；该点是否面向灯光或处于阴影中；以及从该点反射的光量有关。大多数表面通过漫反射过程反射光。漫反射在离开曲面的方向上均匀散射光，因此漫反射曲面的亮度不取决于观察方向。而镜面反射使入射光离开曲面的方向由光到达的方向决定，这就是为什么可以在镜子中分辨不同的对象。镜面反射通常出现在金属表面、涂漆表面、塑料表面和潮湿表面上。它们很容易识别，因为它们又小又亮。几乎在所有情况下，只要将所有曲面建模为具有镜面反射的漫反射曲面就足够了。

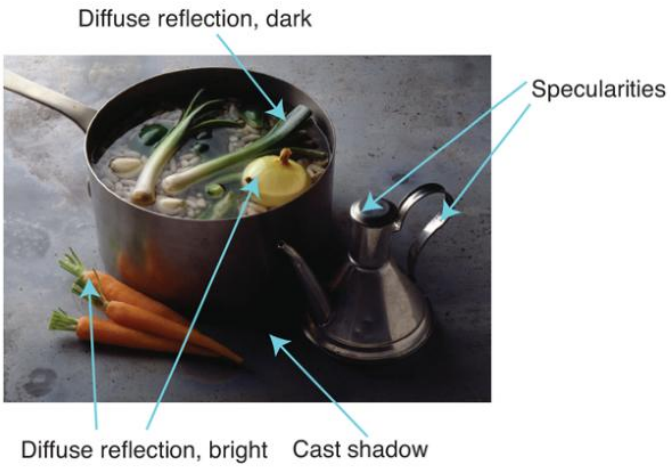


图 2-4 各种照明效果

除了光影，颜色也是视觉形成中的重要因素。托马斯·杨在 1802 年首次提出了这一观点，认为人类观察者只需混合适量的三个原色，就可以匹配任何光谱能量密度的视觉外观。一个常见的选择是红色，绿色和蓝色，缩写为 RGB，这意味着我们可以用每像素 RGB 值的三个数字来表示彩色图像。对于大多数计算机视觉应用，将表面建模为具有三种不同（RGB）漫反射反照率，并将光源建模为具有三种（RGB）强度是足够精确的。

3. 简单的图像属性

图像和视频有四个特别普遍的属性：边缘、纹理、光流和区域分割。

3.1 边缘

边缘是图像平面中的直线或曲线，在其上图像亮度有“显著”变化。边缘检测的目标是从凌乱的、多兆字节的图像中抽象出来，并朝着更紧凑的表示方式发

展。深度不连续、曲面法线发生变化、表面反射率发生变化、照明不连续时都会导致图像强度发生变化，从而出现边缘，如图 3-1。

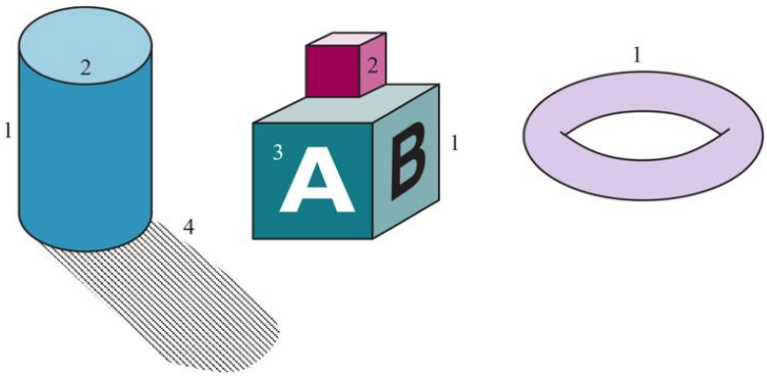


图 3-1 各种边缘效应

然而实际图像中可能存在“噪声”，相机中可能存在热噪声；物体表面可能有划痕，造成表面法线的改变；地表反照率可能会有微小的变化；等等每种效果都会使渐变看起来很大，但不是表示存在边缘。为了避免这种现象，我们通常采取平滑图像，一种常见的方法是将图像通过高斯滤波器，平滑包括使用周围像素来抑制噪声，我们将预测像素的“真实”值，即附近像素的加权和。

3.2 纹理

在计算视觉中，纹理指的是表面上可以被视觉感知的图案。例如建筑物上的窗户、毛衣上的缝线、草坪上的草叶、海滩上的鹅卵石，以及体育场里的人群。纹理表示对于以下两个关键任务非常有用。首先是识别物体——斑马和马的形状相似，但纹理不同。第二个是将一幅图像中的物体与另一幅图像中的同一物体进行匹配，这是从多幅图像中恢复 3D 信息的关键步骤。

3.3 光流

当相机和场景中的一个或多个对象之间存在相对运动时，图像中产生的明显运动称为光流。这描述了由于观察者和场景之间的相对运动，图像中特征的运动方向和速度。例如，从行驶中的汽车上观察到的远处物体的运动比附近物体慢得多，所以光流的速率可以告诉我们距离的一些信息。在图 3-2 中，展示了网球运动员视频中的两帧。在右边显示了根据这些图像计算出的光流矢量。光流对网球运动员移动的场景结构和背景不移动的有用信息进行编码，揭示了运动员正在做的事情：一只手臂和一条腿移动得很快，而其他身体部位没有。

图 3-2 光流矢量图

3.4 自然图像分割

分割是将图像分割成相似像素组的过程。其基本思想是，每个图像像素都可以与某些视觉特性相关联，例如亮度、颜色和纹理。在对象内部或对象的单个部分中，这些属性的变化相对较小，而在对象间边界上，这些属性中的一个或多个通常会发生较大的变化。我们需要找到一个将图像分割成像素集的分区，以便尽可能地满足这些约束。研究这个问题有两种方法，一是侧重于检测这些群体的边界，另一种侧重于检测群体本身，即区域。图 3-3 显示了（b）中的边界检测和（c）和（d）中的区域提取。

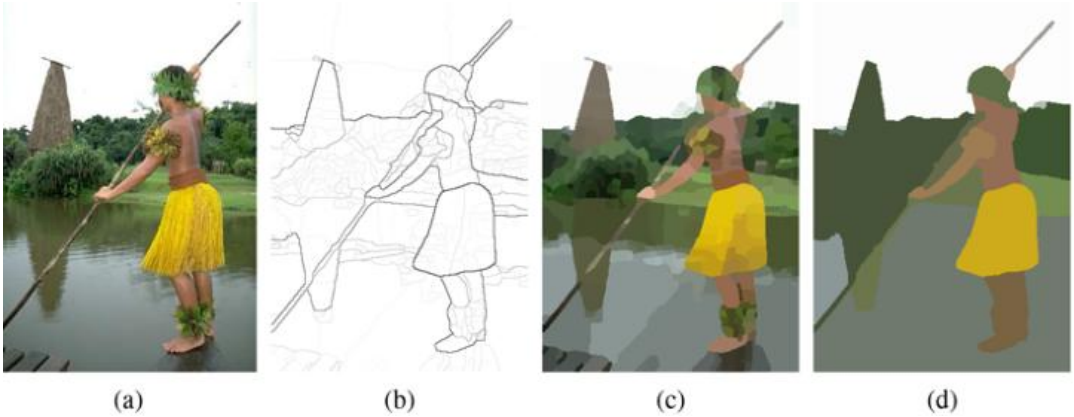


图 3-3 边界检测与区域提取

4. 图像分类

图像分类主要适用于两种情况。第一种，是基于类别分类法的对象，图片中没有太多其他意义。例如，衣服或家具图像，背景无关紧要，分类器的输出是“羊绒衫”或“桌椅”。在另一种情况下，每个图像显示一个包含多个对象的场景。在草原上，你可能会看到长颈鹿和狮子，在客厅里，你可能会看到沙发和台灯，但你不会在客厅里看到长颈鹿或狮子。有了大规模图像分类的方法，可以准确地

输出“草地”或“客厅”。然而同一种类物体看上去颜色、纹理可能不同，在不同时间观测的同一物理也可能由于光线、遮挡、观测角度的不同而不同，如图 4-1。

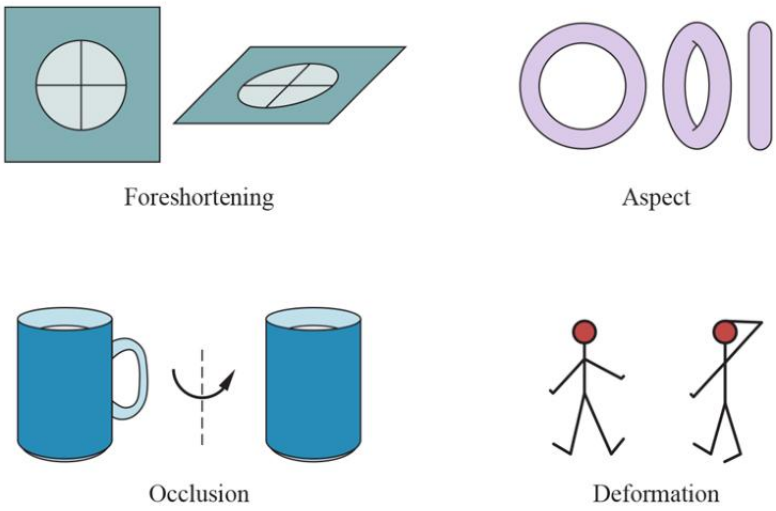


图 4-1 几种效果对图像分类的影响

卷积神经网络（CNN）是非常成功的图像分类器。只要有足够的训练数据和足够的训练能力，就可以产生非常成功的分类系统，这比任何其他方法产生的分类系统都要好得多。Image-Net 数据集在其中发挥了历史性的作用，它为图像分类系统提供了 1400 多万张训练图像，这些图像分为 30000 多个细粒度类别。CNN 分类器使用的功能是从数据集中学习的，而不是由研究人员手工制作的，这确保了这些特征实际上对分类有用。

除了 Image-Net 数据集外，MNIST 数据集是一种常用的标准热身数据集，它包含 70000 张手写数字图像的集合。

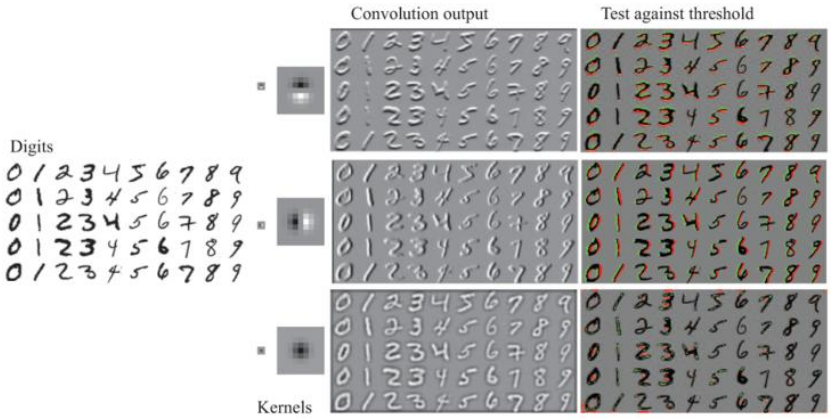


图 4-2 MNIST 数据集

图像的一个重要特性是局部模式可以提供大量信息。例如，数字 0、6、8 和 9 有循环；数字 4 和 8 有交叉点；数字 1、2、3、5 和 7 有行尾，但没有循环或交叉；数字 6 和 9 有循环和行尾。此外，局部模式之间的空间关系也提供了信息：7 有一

条线端在上方；6 有一条线在一个环的上方结束。这些观察结果表明，构建的特征能够响应小型局部社区的模式，然后其他特征再查看这些特征的模式，以此类推。这就是多核卷积网络所擅长的，即网络在多个层次上创建模式，并通过从数据中学习而不是让程序员给出模式来实现。

当我们训练 CNN 时，一个比较实用的技巧是数据集扩充，在这个过程中，对训练示例进行复制和轻微修改。例如，随机移动、旋转或拉伸少量图像，或者随机移动少量像素的色调。在数据集中引入视点或照明的模拟变化有助于增加数据集的大小，尽管新示例与原始示例高度相关，也可以在测试时使用增强功能，而不是在训练时使用。在这种方法中，图像被复制和修改多次（例如，使用随机裁剪），并且分类器在每个修改后的图像上运行，然后使用每个副本的分类器输出对整个类的最终决策进行投票。

5.目标检测

图像分类器预测图像中的内容，并将整个图像分类为多个类别。而对象检测器在图像中查找多个对象，报告每个对象的类别，并通过在对象周围提供边界框来报告每个对象的位置。我们可以通过在更大的图像上观察一个小的滑动窗口来构建一个物体检测器。在每个点，我们使用 CNN 分类器对窗口中看到的内容进行分类。在像素图像中搜索所有可能的窗口是不高效的，但我们知道，包含对象的窗口往往具有相当一致的颜色和纹理。另一方面，将对象一分为二的窗口需要穿过窗口侧面的区域或边。因此，有一种机制可以评分“对象性”——不管一个框里是否有对象，或是什么对象。我们可以找到那些看起来像里面有物体的框，然后对那些通过对象测试的窗口中的物体进行分类。

查找包含对象的区域的网络称为区域建议网络（RPN）。它构建一个网络，以预测每个盒子的分数，并训练这个网络，使盒子包含对象时分数大，否则分数小。根据神经网络结构构造一个 3D 块,块中的每个空间位置都有两个维度作为中心点，一个维度作为盒子类型。任何具有足够好的对象性分数的框都被称为感兴趣区域（ROI），必须由分类器进行检查。但是 CNN 的分类器更喜欢固定大小的图像，而通过对象性测试的盒子在大小和形状上会有所不同。我们不能让盒子拥有相同数量的像素，但我们可以通过采样像素来提取特征，让它们拥有相同数量的特征，这一过程称为 ROI pooling。然后将该固定大小的特征映射传递给分类器。

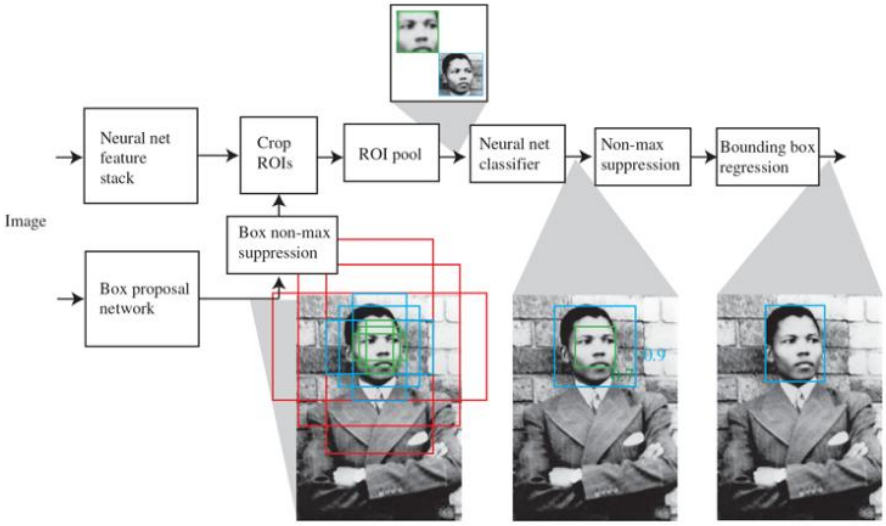


图 5-1 目标检测过程

如图 5-1 所示，一张照片被送入物体探测器，第一个网络计算以网格点为中心的候选图像框的分数。对于示例图像，内部的绿色框和外部的蓝色框通过了对象性测试。第二个网络是一个特征堆栈，用于计算适合分类的图像表示。对象性得分最高的框从特征图中切割出来，通过 ROI pooling 进行大小标准化，并传递给分类器。蓝色框的得分高于绿色框，并且与绿色框重叠，因此绿色框被拒绝。最后，边界框回归蓝色框，使其适合面部。

6. 3D 世界

平面图片中包含了关于三维世界的丰富信息。比如，当我们有关于同一三维空间的多张图片时，可以在图片之间匹配相同点；或者在一张图片中获得其他物体的线索等。下面是几种常见的关于三维世界信息的获取方法。

6.1 多个视图的提示

如果你从不同的视角拍摄了同一场景的两幅图像，那么你可以通过计算出第一个视图中的哪个点对应于第二个视图中的哪个点，并应用一些几何图形，来构建一个 3D 模型。如果有足够多点的两个视图，并且知道第一个视图中的哪个点对应于第二个视图中的哪个点，则无需了解太多有关相机的信息即可构建三维模型。两个点的两个视图提供了四个坐标，只需三个坐标即可在三维空间中确定一个点，额外的坐标则有助于了解关于摄像机的信息。关键问题是如何确定第一个视图中的哪个点对应于第二个视图中的哪个点，我们可以使用简单纹理特征对点的局部外观进行详细描述进行匹配点。

6.2 双目立体视觉

大多数脊椎动物都有两只眼睛，使他们能够使用双目立体视觉以扩大视野。将两个食指举在脸前，一只眼睛闭着，并调整它们，使前指在睁眼的视野中挡住另一只手指。然后交换

左右眼，手指之间的位置发生了变化。这种从左视图到右视图的位置变化称为视差，如果我们在某个深度处叠加一个物体的左右图像，该物体在叠加图像中水平移动，移动的大小是深度的倒数。如图 6-1 中金字塔的最近点在右图中向左移动，在左图中向右移动。

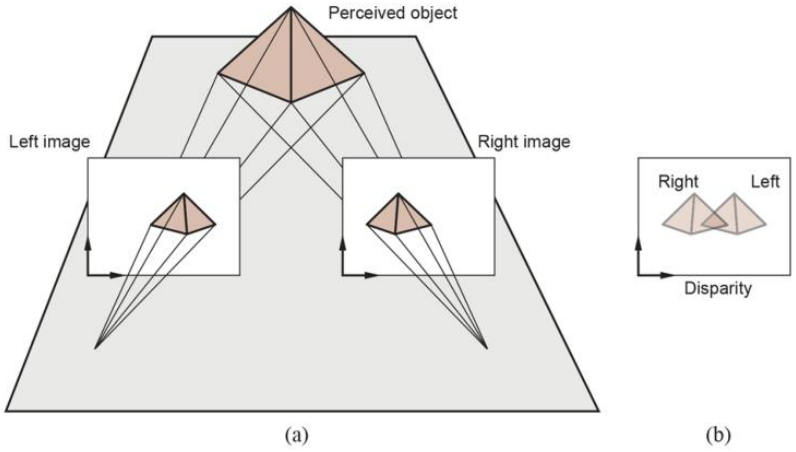


图 6-1 左右视角图

6.3 移动相机

假设我们有一个摄像机在场景中移动，不同图像对应不同时间属性。由于几何结构没有改变，所以立体视觉讨论中的所有模型在相机移动时也适用。我们在这里所说的视差，则是图像中的视在运动，被称为光流。它是摄影机移动和场景几何体的信息源，与观看者的平移速度和场景中的深度相关。

假设你是一只试图降落在墙上的苍蝇，你想从光流场中获得有用的信息。它无法告诉你到墙的距离或到墙的速度，因为尺度模糊。但是如果把距离除以速度，模糊就会消失，得到接触时间，这对于控制着陆非常有用。再考虑两个不同深度的点。我们可能不知道其中任何一个的绝对值，但通过考虑这两点光流大小之比的倒数，我们可以确定深度比。这是运动视差的提示，当我们从移动的汽车或火车的侧窗向外看时，可以推断出风景中移动较慢的部分距离较远。

6.4 单图像

即使是一张图像也能提供有关 3D 世界的丰富信息，即使图像只是一个线条图。因为人们对 3D 形状和布局有一种感觉，尽管这幅画包含的信息很少。遮挡也是一个关键的信息来源，如果图片中有证据表明一个物体遮挡了另一个物体，那么遮挡的物体离眼睛更近。有很好的证据表明，遮挡是构建 3D 结构的有力线索，因为场景中曲面不同部分接收的光强度的变化由场景的几何体和曲面的反射特性决定。

另外，如果图片中有一个物体，那么它的形状很大程度上取决于它的姿势，也就是它相对于观看者的位置和方向。恢复已知对象的姿势有很多应用。例如，在工业操作任务中，机器人手臂在姿势已知之前无法拾取对象；机器人手术应用依赖于精确计算摄像机位置与手术工具和患者位置之间的转换。

最后，物体之间的空间关系是另一个重要线索。下面是一个例子：所有的行人都差不多高，他们站在地平面上。如果我们知道地平线在图像中的位置，我们就可以根据与摄像机的距离对行人身高进行排名。这是因为我们知道他们的脚在哪里，在图像中脚离地平线越近的行人离摄像机越远，因此在图像中越小。反过来，如果场景中有几个行人与摄像机的距离不同，那么一个合理可靠的行人检测器能够估计出地平线，这是因为行人的相对比例也反推出地平线的位置。

7. 计算机视觉应用

计算机视觉的应用范围十分广阔，包括无人驾驶、无人机、目标检测、目标识别、医学图像处理、VR/AR 等，其基本原理与过程和前文所说别无二致，主体就是分类与识别两部分。下面将介绍一些在日常生活场景中计算机视觉的应用过程。

7.1 行为预测

如果我们能够通过分析视频或图像来理解人们在做什么，我们就可以构建观察人们并对他们的行为做出反应的系统。这样我们可以通过收集和使用人们在公共场合的行为数据，更好地设计建筑物和公共场所；建立更准确、侵入性更小的安全监控系统；建立自动化的体育评论员；在人员和机器危险靠近时发出警告，使建筑工地和工作场所更加安全；制作能让玩家站起来走动的电脑游戏；通过管理建筑物内的热量和光线，以匹配居住者所在的位置和他们正在做的事情，从而节约能源.....

然而如何将人们的行为分类以及如何通过对身体和附近物体的观察推测人的目标意图是困难的，因为有时相似的行为看起来不同而不同的行为可能看起来相同，如图 7-1

图 7-1 开冰箱示意图

另一个困难是时间尺度造成的，如图 7-2 所示，一个人在做什么很大程度上取决于时间尺度。该图揭示的另一个重要现象是，多个已识别的行为组成，可以组合成一个更高级别的行为，例如固定零食。

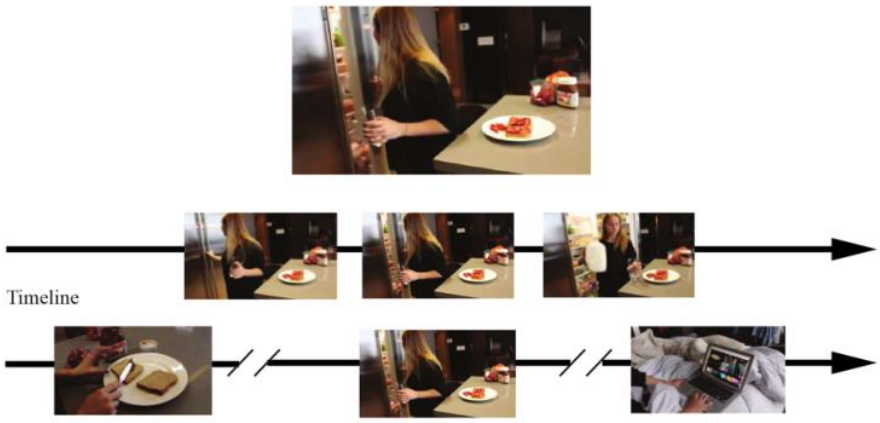


图 7-2 行为合成

事实证明，当训练数据和测试数据来自同一分布时，我们观察到图像分类器和对象检测器工作得非常好。但对于活动数据，训练和测试数据之间的关系不可信，因为人们在如此多的环境中做了如此多的事情。例如，假设有一个行人检测器，它在大数据集上表现良好，训练集中不会出现罕见的现象（例如，骑独轮车），因此我们无法确定探测器在这种情况下会如何工作。挑战在于证明无论行人做什么，探测器都必须是安全的，就像在无人驾驶中，无论行驶的路况多么错综复杂，都要保证乘客生命安全不受威胁。

7. 2 图文标记

互联网上有许多图片和视频。通常，人们希望使用文字，而不是草图进行搜索。因为大多数图片都没有附带文字，所以我们很自然地尝试建立标记系统，用相关文字标记图片，这也是搜索引擎所做的工作。机制很简单，我们应用图像分类和目标检测法，并用输出的文字标记图像。但标记并不能全面描述图像中发生的事情，例如，用对象类别“猫”、“街道”、“垃圾桶”和“鱼骨”标记街道上一只猫的图片时，会忽略猫正在将鱼骨从街道上一个打开的垃圾桶中拉出的信息。 自然而然，我们想到了用句子来描述图片，即字幕系统。其底层机制同样是直接将卷积网络（图像）耦合到递归神经网络或变换网络（句子），并使用标题图像的数据集训练结果。

当前的字幕标注方法使用检测器来查找描述图像的一组单词，并将这些单词提供给经过训练生成句子的序列模型。最准确的方法是搜索模型可以生成的句子，用一组分数评估句子，以找到最好的。强化学习方法可用于训练获得非常好分数的网络。通常训练集中会有一个图

像，其描述与测试集中的图像具有相同的词集；在这种情况下，字幕系统只需检索有效的字幕，而不必生成新的字幕。

图 7-2 强化学习

字幕系统在面对它们无法理解的细节会使用上下文线索猜测。例如，字幕系统往往不善于识别图像中人物的性别，所以经常根据训练数据统计进行猜测。这可能会导致错误——有男人喜欢购物，女人也可能喜欢滑雪。确定一个系统是否能很好地反映图像中发生的事情的一种方法是强迫它回答有关图像的问题。另一种选择是可视化对话系统，它提供了一张图片、标题和一个对话。然后，系统必须回答对话框中的最后一个问题，如图 7-3 所示，但该系统经常出错。

图 7-3 可视化 VQA 系统

7.3 3D 重构

在前面，我们已经讲述了如何通过提取二维图片中的信息来建立三维世界，包括对应点之间的匹配与重建，其实际应用有：1、模型构建：例如，构建一个建模系统，该系统可以获取许多描述对象的视图，并生成纹理多边形的非常详细的 3D 网格，用于计算机图形学和虚拟现实应用程序。2、将动画与真人演员混合：为了将计算机图形角色放入真实视频中，我们需要知道摄影机在真实视频中的移动方式，以便正确渲染角色，并在摄影机移动时更改视图。3、路径重建：如果移动机器人有摄像头，我们可以建立摄像头穿越世界的路径模型，

这将作为机器人路径的表示。4、施工管理：建筑物是极其复杂的人工制品，在施工过程中跟踪发生的事情既费时又费力。保持跟踪的一种方法是每周驾驶无人机穿越施工现场一次，拍摄当前状态。然后三维模型，并使用可视化技术探索计划与重建之间的差异，如图 7-4。

图 7-4 建筑工地 3D 重构图

7.4 视觉操纵

视觉的一个主要用途是提供操纵物体的信息——捡起、抓住、旋转物体等等，以及在避开障碍物的同时导航的信息。假设我们想要建造一辆自动驾驶汽车，其感知系统必须支持以下任务：1、横向控制：确保车辆安全地保持在车道内，或在需要时平稳地变换车道。2、纵向控制：确保与前方车辆保持安全距离。3、避障：监控相邻车道上的车辆，并做好躲避机动的准备。检测行人，让他们安全地通过。4、遵守交通信号：包括交通信号灯、停车标志、限速标志和警察手势。对于横向控制，驾驶员（人类或计算机）需要保持车辆相对于车道的位置和方向表示。对于纵向控制，驾驶员需要与前面的车辆保持安全距离。避障和跟随交通信号需要额外的推断。

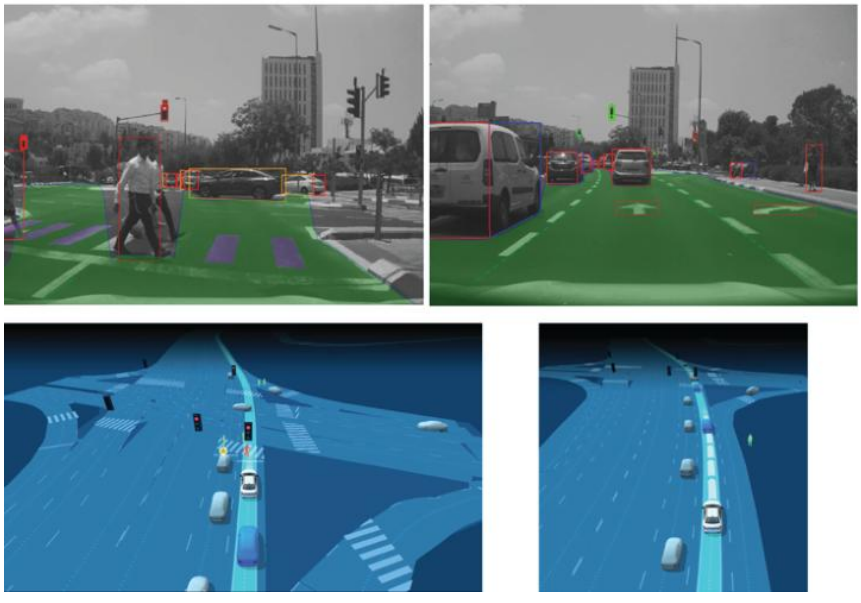


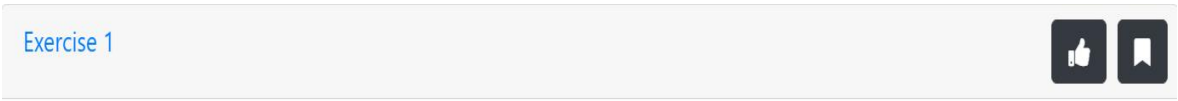
图 7-5 基于 Mobile-Eye 的自动驾驶传感系统

道路是为使用视觉导航的人设计的，因此原则上应该可以单独使用视觉驾驶。然而，实际上，商用自动驾驶汽车使用各种传感器，包括摄像头、激光雷达、雷达和麦克风。激光雷达或雷达可以直接测量深度，这比仅使用视觉的方法更精确。拥有多个传感器通常会提高性能，在能见度低的情况下尤为重要；例如，雷达可以穿透阻挡摄像头和激光雷达的雾气。麦克风可以在接近的车辆（尤其是带有警报器的车辆）变得可见之前检测到它们。

8. 总结

通过本次阅读，我初步了解了计算机视觉的有关知识，掌握了图像如何形成、基本属性、图像分类、目标检测与识别、3D 世界的重建以及一种非常重要的工具：卷积神经网络 CNN，这也为我以后研究模式识别方向打下了坚实的基础。

Quiz：



In the shadow of a tree with a dense, leafy canopy, one sees a number of light spots. Surprisingly, they all appear to be circular. Why? After all, the gaps between the leaves through which the sun shines are not likely to be circular.

Answer:

浓密的枝叶间的空隙很小，这些空隙可以看成是小孔，由于太阳光直线传播，通过小孔照在地面上形成光斑，这些圆形小光斑实际上是太阳通过枝叶间的小孔成像。

参考文献:

《Artificial Intelligence A Modern Approach Fourth Edition》— Chapter25 Computer Vision