计算机视觉与图像识别

Computer Vision and Image Recognition

张国云　郭龙源　吴健辉　胡文静　著

科学出版社

计算机视觉与图像识别

张国云　郭龙源　吴健辉　胡文静　著

科学出版社

北京

前　言

计算机视觉和图像识别是具有很大发展前景的前沿研究领域。计算机视觉领域的突出特点是其多样性与不完善性。尽管人们已开始掌握部分解决具体计算机视觉任务的方法，但这些方法通常都仅适用于一群狭隘的目标（如脸孔、指纹、文字等），因而无法被广泛地应用于不同场合。对这些方法的应用通常作为某些解决复杂问题的大规模系统的一个组成部分（例如医学图像的处理、工业制造中的质量控制与测量）。正是因为这些原因，计算机视觉技术一直吸引着众多学者持续的研究兴趣。

图像识别技术是通过计算机，采用数学方法，对系统获取的图像按照特定目的进行相应的处理。常见的如生物特征识别（人脸识别、指纹识别等）、智能交通中的目标识别、工业中的工件识别等。可以说，图像识别技术是人类视觉认知的延伸，是人工智能的一个重要领域。顺理成章地，对图像进行识别也是计算机视觉的重要任务。相信随着计算机技术及人工智能技术的发展，它们涉及的技术领域会越来越广泛，应用也会越来越深入。

本书结合机器视觉和图像识别领域的基本理论，以湖南理工学院图像信息处理与智能系统科研团队中的四位博士论文为基础，结合近年团队成员科研成果和所发表的相关学术论文，系统阐述了计算机视觉理论和立体匹配算法，支持向量机、人脸识别等基本理论和技术方法的应用。

本书是作者在湖南省自然科学基金项目（06JJ50133）、湖南省教育厅重点项目（10A046）、湖南省教育厅优秀青年项目（05B052）、湖南省教育厅项目（09C471）、上海市应用材料科技国际共同计划（上海市 AM 基金）项目和国防基础研究项目（J1500C002）的资助下科学研究与应用工作的积累。主要内容包括计算机视觉和支持向量机的基本原理、三种立体匹配算法的研究、人脸检测识别算法的研究、基于视觉的障碍物检测算法研究等。按计算机视觉立体匹配，支持向量机在人脸检测识别中应用等不同研究方向，分别独立成章。

书中提及的理论与技术可以广泛应用于工业、军事及民用中，对推动计算机视觉和图像识别技术的应用具有很好的价值，可作为计算机视觉、模式识别、人工智能、图像处理等有关专业的研究工作者的参考用书。

目　录

第1章 绪 论

视觉是人类和其他一些动物所具有的基本功能。人们通过眼睛来观察周围的客观世界，识别环境物体，辨明物体的方位，确定周围事物的组成及相互之间的关系。人们也可以通过视觉来识别文字图像等资料，理解其中的含义。人类所具有的这种特有的功能称为视觉功能。这是人们认识世界、了解世界的重要途径之一。视觉，不仅指对光信号的感受，也包括对视觉信息的获取、传输、处理、存储与理解的全过程。信号处理理论与计算机出现以后，人们试图用摄像机获取环境图像并转换成数字信号，并且用计算机实现对视觉信息处理的全过程，这样就形成了一门新的学科——计算机视觉。现在该学科在许多领域都得到了广泛的应用。

1.1 计算机视觉的目标与任务

在自然界中存在着各种各样的物体，而人们则是通过“看”来了解这些物体的。如果视觉指的是通过最终手段——“看”来了解世界，那么计算机视觉的应用研究也是追求这同一目的，即通过图像对客观物体建立明确而有意义的描述。具体来说，就是让计算机具有对周围世界的空间物体进行传感、抽象、判断的能力，从而达到识别、理解的目的。根据其处理过程的先后及复杂程度，计算机视觉的任务可以分为几个方面。

1. 图像获取

图像获取是指通过光学摄像机、红外摄像机或激光、超声波、雷达等对周围视觉世界进行传感，使计算机得到与视觉世界相对应的二维图像。该图像由数字组成，常称为数字图像。获取数字图像的功能是计算机视觉系统应具有的必要的基本功能。

2. 特征抽取

物体的表面色彩、纹理或轮廓形状，统称为特征或特性。特征是用来区别一个视觉物体与另一个物体的重要根据。特征抽取就是指在二维图像基础上，分析物体的特征。特征抽取的结果是参数、表格等表达方式。

3. 识别与分类

识别与分类是指根据预先在计算机中的物体的“模式”，对特征抽取的结果进行分析、比较，判定图像中感兴趣的物体是否存在并确定它的位置，或区分、标识图像中各物体的类别。

4. 三维信息理解

三维信息理解是指根据已知物体的三维模型从二维图像中估计推断出物体的三维立体信息，包括物体的三维空间位置、表面形状的朝向等。

5. 景物描述

景物描述是指分析二维图像中各种物体的结构及它们的相互关系，或推断对应的三维空间中各物体组成及三维空间关系。前者常称为图像的描述，后者称为三维景物的描述。

6. 图像解释

有可能的话，除了景物描述外还得指出视觉世界以外的含义，例如从景物的现状说明其中的含义、原因及下一步将会发生什么等。这个过程称为图像解释。

1.2 计算机视觉的经典问题

在每个计算机视觉技术的具体应用中，人们都要解决一系列相同的问题，这些经典的问题包括以下几个方面。

1. 识别

计算机视觉和图像处理所共有的经典问题是判定一组图像数据中是否包含某个特定物体的图像特征或运动状态，如简单几何图形识别、人脸识别、印刷或手写文件识别及车辆识别。这些识别需要在特定的环境中，具有指定的光照、背景和目标姿态要求。

广义的识别在不同的场合又演化成了以下几个略有差异的概念。

① 识别（狭义的）。对一个或多个经过预先定义或学习的物体、物类进行辨识，通常在辨识过程中还要提供其二维位置或三维姿态。

② 鉴别。识别、辨认单一物体本身，如某一人脸的识别、某一指纹的识别等。

③ 监测。从图像中发现特定的情况内容，如医学中对细胞或组织不正常技能的发现、交通监视仪器对过往车辆的发现等。监测往往是通过简单的图像处理发现图像中的特殊区域，为后继更复杂的操作提供起点。

2. 运动检测

运动检测是指在指定区域能识别图像的变化，检测运动物体的存在并避免由光线变化带来的干扰。为了从实时的序列图像中将变化区域从背景图像中提取出来，需要将运动区域进行有效分割，其对于目标分类、跟踪等后期处理是非常重要的，因为以后的处理过程仅仅考虑图像中对应于运动区域的像素。然而，由于背景图像的动态变化，如天气、光照、影子及混乱干扰等的影响，使得运动检测成为一项相当困难的

工作。

基于序列图像对物体运动的检测包含多种类型，如自体运动（监测摄像机的三维刚性运动）和图像跟踪（跟踪运动的物体）。目前常用的方法有背景减除法（Background Subtraction）、时间差分法（Temporal Difference）和光流法（Optical Flow）。还有一种检测方法称为运动向量检测法，其适合于多维变化的环境，能消除背景中的振动像素，使某一方向的运动对象更加突出地显示出来。但运动向量检测法也不能精确地分割出对象。

3. 场景重建

给定一个场景的多幅图像或者一段录像，通过场景重建寻求为该场景建立一个计算机模型（三维模型）。最简单的情况便是生成一组三维空间中的点，更复杂的情况下会建立起完整的三维表面模型。

目前，获取三维信息的方法和技术很多，学术思想非常活跃，新技术、新方法不断涌现。双目立体视觉是计算机视觉被动测距方法中最重要的距离感知技术，它直接模拟了人类视觉处理景物的方式，可以在多种条件下灵活地测量景物的立体信息，其作用是其他计算机视觉方法所不能取代的。对它的研究，无论是从视觉生理的角度还是在工程应用中都具有十分重要的意义。

4. 图像恢复

图像恢复的目标在于移除图像中的噪声，如仪器噪声、模糊等。具体来讲，是通过计算机处理，对质量下降的图像加以重建或恢复的处理过程。因摄像机与物体相对运动、系统误差、畸变、噪声等因素的影响，图像往往不是真实景物的完善映像。图像恢复就是建立造成图像质量下降的退化模型，然后运用相反过程来恢复原来图像，并运用一定准则来判定是否得到图像的最佳恢复。例如在遥感图像处理中，为消除遥感图像的失真、畸变，恢复目标的反射波谱特性和正确的几何位置，通常需要对图像进行恢复处理，包括辐射校正、大气校正、条带噪声消除、几何校正等内容。

接下来本书将对双目立体视觉和图像识别等问题展开深入研究。

1.3 Marr的计算机视觉理论框架

20世纪80年代初，Marr首次从信息处理的角度综合了图像处理、心理物理学、神经生理学及临床精神病学的研究成果，提出了第一个较为完善的视觉系统框架，这一框架虽然在细节甚至在主导思想方面尚存在大量不完备之处，许多方面还有许多争议，但至今仍是广大计算机视觉工作者接受的基本框架，计算机视觉这门学科的形成，应该说与这一理论框架有密切的关系。下面从几个方面来描述这一理论框架。

1.3.1 视觉系统研究的三个层次

Marr从信息处理系统的角度出发，认为对此系统的研究应分为三个层次，即计算

理论层次、表达与算法层次、硬件实现层次。

计算理论层次要回答系统各个部分的计算目的与计算策略，即各部分的输入/输出是什么，之间的关系是什么。Marr 对视觉系统的总的输入/输出关系规定了一个总的目标，即输入是二维图像，输出是由二维图像重建出来的三维物体的位置与形状。Marr 认为，视觉系统的任务是对环境中三维物体进行识别、定位与运动分析，但这仅仅是一种对视觉行为的目的性定义，而不是从计算理论层次上的目的性定义。三维物体千差万别，应存在一种计算层次上的一般性目的描述。达到了这个目的，则不管是什么具体的物体，视觉任务均可完成。Marr 认为这一目的就是要通过视觉系统，重建二维物体的形状、位置。而且如果在每一时刻都能做到这一点，则运动分析也可以做到。对视觉系统的各个层次与模块，Marr 也初步给出了计算理论层次上的目标。

对于表达与算法层次，视觉系统的研究应给出各部分（或称各模块）的输入、输出和内部的信息表达，以及实现计算理论所规定的目标的算法，算法与表达有关。不同的表达方式，完成同一计算的算法会不同。但 Marr 认为，算法与表达是比计算理论低一层次的问题，不同的表达与算法在计算理论层次上可以是相同的。

硬件层次是要问答“如何用硬件实现以上算法”。

从信息处理的观点来看，至关重要的是最高层次，即计算理论层次。这是因为构成知觉的计算本质，取决于解决计算问题的本身，而不取决于用来解决计算问题的特殊硬件。通过正确理解待解决问题的本质，将有助于理解并创造算法。如果仅考虑解决问题的机制和物理实现，则对理解算法往往无济于事。

区别以上三个不同层次，对于深刻理解计算机视觉与生物视觉系统以及它们的关系都是有益的。例如，人的视觉系统与目前的计算机视觉系统在硬件实现层次上是完全不同的，前者是极为复杂的神经网络，而后者是目前使用的计算机。但它们可能在计算理论层次上完成相同的功能。视觉系统研究的这三个层次的含义总结为表 1.1。

表 1.1 视觉系统研究的三个层次的含义

名　称	含义和所解决的问题
计算理论	计算的目的
表达与算法	实现计算理论的方法，输入/输出的表达
硬件实现	如何在物理上实现表达和算法

1.3.2 视觉信息处理的三个阶段

Marr 从视觉计算理论出发，将系统分为自下而上的三个阶段，即视觉信息从最初的原始数据（二维图像数据）到最终对三维环境的表达经历了三个阶段的处理。第一阶段（早期阶段）构成所谓“要素图”或“基元图”，基元图由二维图像中的边缘点、直线段、曲线、顶点、纹理等基本几何元素或特征组成；第二阶段称为对环境的 2.5 维描述。2.5 维描述是一种形象的说法，意即部分的、不完整的三维信息描述，用“计算”的语言来讲，就是重建三维物体在观察者为中心的坐标系下的三维形状与位置。当人眼或摄像机观察周围环境物体时，观察者对三维物体最初是以自身的坐标系来描

述的。另外，我们只能观察到物体的一部分（另一部分是物体的背面或被其他物体遮挡的部分）。这样，重建的结果是以观察者坐标系下描述的部分三维物体形状，称为2.5维描述。这一阶段中存在许多并行的相对独立的模块，如立体视觉、运动分析、由灰度恢复表面形状等不同处理单元。事实上，从各种不同角度去观察物体，观察到的形状都是不完整的，不能设想，人脑中存有同一物体从所有可能的观察角度看到的物体形象，以用来与所谓的物体的2.5维描述进行匹配与比较。因此2.5维描述必须进一步处理以得到物体完整的三维描述，而且必须是物体本身某一固定坐标系下的描述，这一阶段称为第三阶段（后期阶段）。视觉信息处理的三个阶段如表1.2所示。

表1.2 由图像恢复形状信息的表达框架

名 称	目 的	基 元
图像	亮度表示	图像中每点的亮度值
基元图	表示二维图像中的重要信息，主要是图像中的亮度变化位置及其几何分布和组织结构	零交叉，斑点，端点和不连续点，边缘，有效线段，组合群，曲线组织，边界
2.5维图	在以观测者为中心的坐标中，表示可见表面的方向、深度值和不连续的轮廓	局部表面朝向离观测者的距离深度上的不连续点
3维模型表示	在以物体为中心的坐标系中，有由体积基元和面积基元构成的模块化多层次表示，描述形状及其空间组织形式	分层次组成若干三维模型，每个三维模型都是在几个轴线空间的基础上构成的，所有体积基元或面积形状基元都附着在轴线上

Marr理论是计算机视觉研究领域的划时代成就，对图像理解和计算机视觉的研究发展起了重要的作用。但Marr理论也有其不足之处，其中有4个有关整体框架的问题（见图1.1）。

（1）框架中输入是被动的，给什么图像，系统就处理什么图像。

（2）框架中加工目的不变，总是恢复场景中物体的位置和形状等。

（3）框架缺乏高层知识的指导作用。

（4）整个框架中信息加工过程基本自下而上，单向流动，没有反馈。

针对上述问题，近年来人们提出了一系列改进思路，对应图1.1的框架，可将其改进并融入新的模块得到图1.2的框架，具体改进如下。

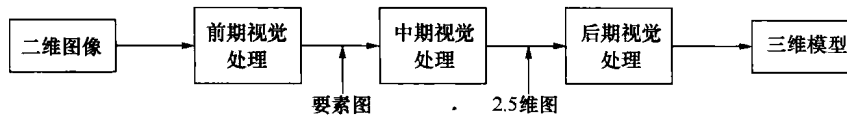

图1.1 Marr的视觉理论框架

人类视觉是主动的，会根据需要改变视角以帮助识别。主动视觉指视觉系统可以根据已有的分析结果和视觉的当前要求，决定摄像机的运动以从合适的视角获取相应的图像。人类的视觉又是有选择的，可以注视（以较高分辨率观察感兴趣区域），也可

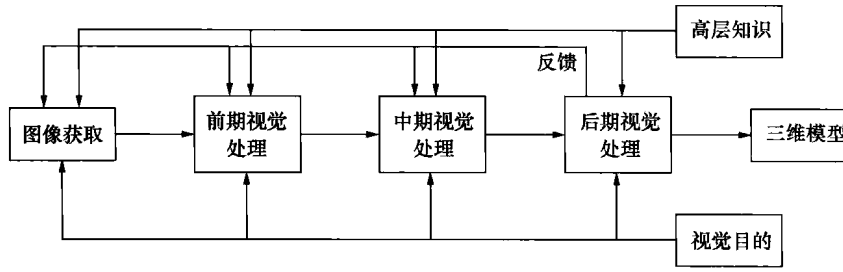

图 1.2 Marr 的视觉理论框架

以对场景中某些部分视而不见。选择性视觉指视觉系统可以根据已有的分析结果和视觉的当前要求，决定摄像机的注意点以获取相应的图像。考虑到这些因素，在改进框架中增加了图像获取模块。该模块要根据视觉目的选择采集方式。

人类的视觉可以根据不同的目的进行调整。有目的视觉（定性视觉）指视觉系统根据视觉目的进行决策。事实上，有相当场合只需定性结果就可以，并不需要复杂性高的定量结果。因此在改进框架中增加了视觉目的模块，但定性分析还缺乏完备的数学工具。

还有一种相关的观点认为 Marr 关于对场景先重建再解释的思路可以简化视觉任务，但与人的视觉功能并不完全吻合。事实上重建和解释不总是串行的。

人类可在利用图像获取部分信息的情况下完全解决视觉问题，原因是隐含地使用了各种知识。例如，借助 CAD 设计资料获取物体形状信息（使用物体模型库），可帮助解决由单幅图恢复物体形状的困难。利用高层知识可解决低层信息不足的问题，所以在改进框架中增加了高层知识模块。

人类视觉中前后处理之间是交互作用的，尽管对这种交互作用的机理了解得还不充分，但高层知识和后期处理的反馈信息对早期处理的作用是重要的。从这个角度出发，在改进框架中增加了反馈控制流向。

最后需要指出，限于历史等因素，Marr 没有研究如何用数学方法严格地描述视觉信息的问题，虽然较充分地研究了早期视觉，但基本没有论及对视觉知识的表达、使用和基于视觉知识的识别等。近年来有许多试图建立计算机视觉理论框架的工作，其中 Grossberg 宣称建立了一个新的视觉理论——表观动态几何学（dynamic geometry of surface form and appearance）。它指出感知的表面形状是分布在多个空间尺度上多种处理动作的总结果，2.5 维图并不存在，向 Marr 理论提出了挑战。但 Marr 的理论使得人们对视觉信息的研究有了明确的内容和较完整的基本体系，仍被看作是研究的主流。现在提出的理论框架均包含它的基本成分，多数被看作是它的补充和发展。尽管 Marr 的理论在许多方面还存在争议，但至今它仍是计算机视觉工作者所普遍接受的计算机视觉理论框架。

1.4 摄像机成像几何模型

摄像机的成像过程是从三维空间到二维图像平面的投影。为了研究图像的投影及

其包含的三维信息，必须首先确定摄像机的投影模型，才能确定三维空间点和其在二维图像平面的投影之间的关系。在不同使用条件下，即摄像机和空间物体之间有不同的位置关系时，可以采取不同的摄像机投影模型。透视投影、平行投影和正透视投影是计算机视觉中最常用的三种摄像机模型，其中正透视投影是透视投影的特例，而当摄像机和空间物体之间的距离比空间物体的最大尺寸大很多的情况下，可以用平行投影来简化透视投影模型。这里主要以透视投影模型为研究前提。

在大部分应用环境中可以用理想的针孔模型来近似实际摄像机。针孔模型的几何关系就是透视投影。场景中的物体最终结果是投射到胶片上的模拟图像或计算机里面的数字图像。对场景和其图像的研究涉及不同坐标系之间的变换，主要的坐标系统有如下几种。

① 世界坐标系。也称真实坐标系统，它是客观世界的绝对坐标。一般三维场景利用这个坐标系统来表示。

② 摄像机坐标系。以摄像机为中心制定的坐标系统。

③ 像平面坐标系。即在摄像机内的像平面上的坐标系统。原点在摄像机的光轴上。

④ 计算机图像坐标系。类似于像平面坐标系，对于不同的计算机操作系统，一般只是坐标原点和计量单位的差别。本书的计算机图像坐标系原点设定在图像的左上角，计量单位是像素值。

摄像机坐标系的定义见图 1.3。将坐标系 z (O, i, j, k) 附加到针孔摄像机上，向量 i，j，k 组成右手直角坐标系。原点 O 与针孔重合，而向量 i 和 j 组成一个与图像平面 Π' 平行的向量平面的基，Π' 平面位于沿 k 向量正方向距离针孔 f' 处。通过针孔垂直于 Π' 的线称为光轴。

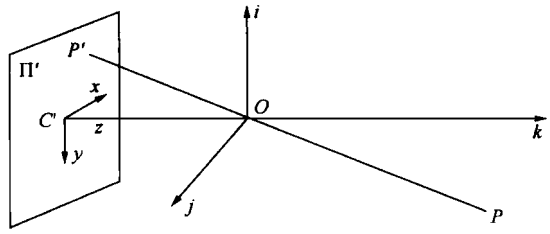

图 1.3 针孔成像模型及摄像机坐标系和像平面坐标系示意图

为表示透视模型还需要在图像平面中建立像平面坐标系。像平面坐标系是一个二维坐标系，在图 1.3 中，光轴穿过 Π' 和图像平面，其交点 C' 称为图像中心，作为像平面坐标系的原点。如果用点 p 表示景物中坐标为 (x, y, z) 的一点，p' 是它的投影像点，坐标为 (x', y', z')。因为 p' 处在图像平面中，所以有 $z'=f'$。又因为 p，O，p' 这三点共线，则应有 $\boldsymbol{OP'}=\lambda\boldsymbol{OP}$，$\lambda$ 为某个数，所以

$$\begin{cases} x' = \lambda x \\ y' = \lambda y \Leftrightarrow \lambda = \dfrac{x'}{x} = \dfrac{y'}{y} = \dfrac{f'}{z} \\ f' = \lambda z \end{cases} \tag{1.1}$$

因此有

$$\begin{cases} x' = f' \dfrac{x}{z} \\ y' = f' \dfrac{y}{z} \end{cases} \tag{1.2}$$

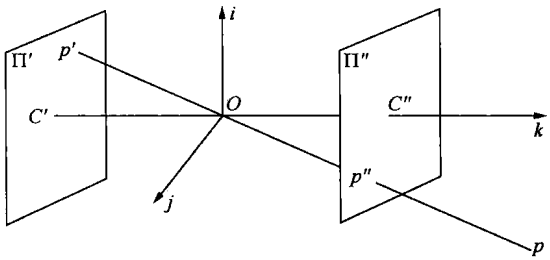

图 1.4 使用虚拟图像平面的针孔成像模型

在实际应用中，物体与摄像机原点的距离一般都远大于焦距。透视投影产生的是一幅颠倒的图像，为了方便，可以设想一个虚拟图像平面。这幅图像落在一个处于焦心前面的平面上，它到焦心的距离等于实际成像面到焦心的距离，即它们关于摄像机原点对称。此虚拟平面被称为像平面，如图 1.4 所示的 Π'' 平面。

1.5 摄像机参数和透视投影

计算机视觉的基本任务之一是从摄像机获取的图像信息出发计算三维空间中物体的几何信息，并由此重建和识别物体，而空间物体表面某点的三维几何位置与其在图像中对应点之间的相互关系是由摄像机成像的几何模型决定的，这些几何模型参数就是摄像机参数。在大多数条件下这些参数必须通过实验与计算才能得到，这个过程被称为摄像机标定。根据摄像机参数性质可以分为内部参数和外部参数：内部参数描述摄像机的内部光学和几何特性（如图像中心、焦距、镜头畸变及其他系统误差参数等）；相对于一个世界坐标系的摄像机坐标的三维位置和方向称为外部参数。这里以简化的摄像机光学成像模型，即针孔模型为基础，讨论透视投影和多个坐标系下的坐标转换关系。

1.5.1 坐标系变换和刚体变换

设有坐标系 F，将该坐标系中点 P 的坐标向量记为${}^{F}\boldsymbol{P}$，即

$$ {}^{F}\boldsymbol{P} = {}^{F}\boldsymbol{OP} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \Leftrightarrow \boldsymbol{OP} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} \tag{1.3} $$

考虑两个坐标系的情况：$(A) = (O_A, i_A, j_A, k_A)$ 和 $(B) = (O_B, i_B, j_B, k_B)$。任务是如何把${}^{B}\boldsymbol{P}$ 表示成${}^{A}\boldsymbol{P}$。

当两个坐标系之间是纯平移关系时，有 $\boldsymbol{O_BP} = \boldsymbol{O_BO_A} + \boldsymbol{O_AP}$，则

$$ {}^{B}\boldsymbol{P} = {}^{A}\boldsymbol{P} + {}^{B}\boldsymbol{O_A} \tag{1.4} $$

当两个坐标系之间是纯旋转关系时，旋转矩阵是个 3×3 的数组，定义为${}^{B}_{A}\boldsymbol{R}$，则

$$ {}^{B}_{A}\boldsymbol{R} \overset{\text{def}}{=} \begin{pmatrix} i_A \cdot i_B & j_A \cdot i_B & k_A \cdot i_B \\ i_A \cdot j_B & j_A \cdot j_B & k_A \cdot j_B \\ i_A \cdot k_B & j_A \cdot k_B & k_A \cdot k_B \end{pmatrix} \tag{1.5} $$

即

$$ {}^{B}_{A}\boldsymbol{R} = ({}^{B}i_A \quad {}^{B}j_A \quad {}^{B}k_A) = \begin{pmatrix} {}^{A}i_B{}^{T} \\ {}^{A}j_B{}^{T} \\ {}^{A}k_B{}^{T} \end{pmatrix} \tag{1.6} $$

其满足${}_A^B\boldsymbol{R}={}_A^B\boldsymbol{R}^T$。

一般来说，旋转矩阵可以分解为绕 i，j，k 旋转的基本旋转矩阵的乘积。由${}_B^D\boldsymbol{R}$ 是单位阵可知，在坐标系 B 中满足

$$ {}^B\boldsymbol{P}={}_A^B\boldsymbol{R}{}^A\boldsymbol{P} \tag{1.7}$$

若两个坐标系的原点和基向量都是不同的，则称这两个坐标系之间是一般的刚体变换，且有

$$ {}^B\boldsymbol{P}={}_A^B\boldsymbol{R}{}^A\boldsymbol{P}+{}^B\boldsymbol{O}_A \tag{1.8}$$

在齐次坐标情况下，上个方程可以写成矩阵乘积的形式，即

$$ \begin{pmatrix}{}^B\boldsymbol{P}\\1\end{pmatrix}={}_A^B\boldsymbol{T}\begin{pmatrix}{}^A\boldsymbol{P}\\1\end{pmatrix} $$

其中

$$ {}_A^B\boldsymbol{T}=\begin{pmatrix}{}_A^B\boldsymbol{R} & {}^B\boldsymbol{O}_A\\ \boldsymbol{O}^T & 1\end{pmatrix},\quad \boldsymbol{O}=(0,0,0)^T \tag{1.9}$$

这样就可以用一个 4×4 矩阵和一个四维向量表示任意的坐标系变换。

1.5.2 摄像机参数和透视投影

在不考虑透镜引起的非线性畸变的前提下，我们讨论摄像机坐标系、像平面坐标系、计算机图像坐标系与世界坐标系之间的投影关系。其中摄像机坐标系和像平面坐标系、计算机图像坐标系之间的关系称为摄像机内参数，而外参数表示摄像机在世界坐标系里的位置和方向。为了得到摄像机的内参数，可以定义一个归一化的图像平面。该平面平行于摄像机的物理成像平面，且到针孔的距离为单位长度。在其上定义的坐标系中，原点定在光轴和这个平面的交点处，即 C'点。则透视投影方程为

$$ \begin{cases}\hat{u}=\dfrac{x}{z}\\ \hat{v}=\dfrac{y}{z}\end{cases}\Leftrightarrow\hat{\boldsymbol{p}}=\frac{1}{z}(\boldsymbol{Id}\quad 0)\begin{pmatrix}\boldsymbol{p}\\1\end{pmatrix} \tag{1.10}$$

其中，$\hat{\boldsymbol{p}}\overset{\text{def}}{=}(\hat{u},\hat{v},1)^T$ 是点 P 投影到这个平面上的 p 点的齐次坐标表示。

像平面坐标中，图像中的点（u，v）一般用像素表示，而不是用米等单位来表示。而且像素一般不是正方形，而是长方形，所以需要用两个额外的比例因子 k 和 l，且

$$ \begin{cases}u=kf\dfrac{x}{z}\\ v=lf\dfrac{y}{z}\end{cases} \tag{1.11}$$

在这里，假设 f 是用米表示的距离，像素的大小是 $1/kl$，其中，k 和 l 的单位是像素/米。参数 k，l 和 f 是相关的，如果用像素单位表示，有 $\alpha=kf$ 和 $\beta=lf$。

而在计算机图像坐标系中，一般把图像的左上角而不是中心定为原点，则需要添加两个参数 u_0 和 v_0 来定义点在计算机图像坐标系中的位置，则式（1.11）改为

$$ \begin{cases}u=\alpha\dfrac{x}{z}+u_0\\ v=\beta\dfrac{y}{z}+v_0\end{cases} \tag{1.12}$$

最后，由于制造误差，摄像机坐标系可能会产生偏离，即两个坐标轴不完全垂直。在这种情况下，式（1.12）修正为

$$\begin{cases} u = \alpha \dfrac{x}{z} - \alpha\cot\theta \dfrac{y}{z} + u_0 \\ v = \dfrac{\beta}{\sin\theta} \dfrac{y}{z} + v_0 \end{cases} \tag{1.13}$$

最终得到

$$\boldsymbol{P} = \frac{1}{z}\boldsymbol{M}\boldsymbol{P}, \quad \text{其中，} \quad \boldsymbol{M} = (\boldsymbol{K} \quad 0), \boldsymbol{K} = \begin{pmatrix} \alpha & -\alpha\cot\theta & u_0 \\ 0 & \dfrac{\beta}{\sin\theta} & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.14}$$

$\boldsymbol{P}=(x, y, z, 1)^T$ 表示摄像机坐标系里的齐次坐标。利用矩阵 $\boldsymbol{M}$，齐次坐标可以表示从四维到三维的投影变换。

摄像机在世界坐标系里的位置和方向是摄像机的外参数。摄像机坐标系 C 在世界坐标系 W 中的位置记为

$$\begin{pmatrix} {}^{C}\boldsymbol{P} \\ 1 \end{pmatrix} = \begin{pmatrix} {}^{C}_{W}\boldsymbol{R} & {}^{C}\boldsymbol{O}_W \\ 0^T & 1 \end{pmatrix} \begin{pmatrix} {}^{W}\boldsymbol{P} \\ 1 \end{pmatrix} \tag{1.15}$$

带入式（1.14）得到

$$\boldsymbol{p} = \frac{1}{z}\boldsymbol{M}\boldsymbol{P} \tag{1.16}$$

其中，$\boldsymbol{M}=\boldsymbol{K}(\boldsymbol{R} \quad \boldsymbol{t})$，$\boldsymbol{K}=\begin{pmatrix} \alpha & -\alpha\cot\theta & u_0 \\ 0 & \dfrac{\beta}{\sin\theta} & v_0 \\ 0 & 0 & 1 \end{pmatrix}$，$\boldsymbol{R}={}^{C}_{W}\boldsymbol{R}$ 是旋转矩阵，$\boldsymbol{t}={}^{C}\boldsymbol{O}_W$ 是平移向量，$\boldsymbol{P}=({}^{W}x, {}^{W}y, {}^{W}z, 1)^T$ 表示向量 $\boldsymbol{P}$ 在坐标系 W 下的齐次坐标。这样，投影矩阵可以表达为

$$\boldsymbol{M} = \begin{pmatrix} \alpha r_1^T - \alpha\cot\theta\, r_2^T + u_0 r_3^T & \alpha t_x - \alpha\cot\theta\, t_y + u_0 t_z \\ \dfrac{\beta}{\sin\theta} r_2^T + v_0 r_3^T & \dfrac{\beta}{\sin\theta} t_y + v_0 t_z \\ r_3^T & t_z \end{pmatrix} \tag{1.17}$$

其中，r_1^T，r_2^T 和 r_3^T 表示 $\boldsymbol{R}$ 的三行，t_x，t_y 和 t_z 是向量 $\boldsymbol{t}$ 的坐标。

第 2 章　立体视觉匹配算法

立体匹配是计算机立体视觉的关键问题。根据匹配基元的不同，立体视觉匹配算法目前分为三大类：区域匹配、特征匹配和相位匹配。如何快速、鲁棒地实现图像对的对应点匹配，获得满足要求的深度图，或称为视差图，是当前研究的热点之一。以下从提高快速性和鲁棒性的思想出发，阐述了三个不同的匹配算法。

2.1　快速区域视差匹配算法

2.1.1　深度信息计算及约束条件

计算机立体视觉是从模仿人的双眼产生立体感知中获得启示，研究如何从左右两个摄像机所得到的两幅视觉图像中获取场景中物体的三维深度（距离）信息，其结果表现为深度图，经过进一步处理就可以得到景物的三维空间信息，实现二维图像到三维空间的重构。

1. 立体成像方式

立体成像的方式主要由光源、采集器和景物三者的相互位置和运动情况所决定。最简单的是单目成像，即用一个采集器在一个固定位置对场景取一幅像。此时有关景物的立体信息是隐含在所成像的几何畸变、明暗度（阴影）、纹理、表面轮廓等因素之中的。如果用两个采集器在不同位置对同一场景取像就是双目成像（对于完全静态场景也可用同一采集器在两个不同位置先后对同一场景取像）。此时两幅像间所产生的视差可用来帮助求取采集器与景物的距离。如果用多于两个的采集器在不同位置对同一场景取像就是多目成像。

以上讨论中认为几种成像方式里光源都是固定的，如果将采集器相对景物固定而将光源绕景物移动就是光移成像（也称立体光度成像）。由于同一景物表面在不同光照情况下亮度不同，所以由光移像可求得物体的表面朝向（但并不能得到绝对的深度信息）。如果保持光源固定而让采集器运动跟踪场景或让采集器和景物同时运动就构成主动视觉成像，其中后一种又称为主动视觉自运动成像。另外如果用可控的光源照射景物，通过采集到的投影模式来解释景物的表面形状就是结构光成像方式。在这种方式中可以将光源和采集器固定而将景物转动，也可以将景物固定而将光源和采集器一起绕景物运动。以上几种方式的一些特点概括在表 2.1 中。

在以上各种方式中，最后在像平面得到的每幅图像都是 2D 的，但这些图像中仍带有原来三维场景中的立体信息（特别是采集器与景物间距离的信息）。