第一章 计算机视觉简介：历史、现状和发展趋势

摘要

对计算机视觉 40 多年的发展历程进行了简要总结，包括：马尔计算视觉理论，主动视觉与目的视觉，多视几何与摄像机自标定，以及基于学习的视觉。在此基础上，对计算机视觉的未来发展趋势给出了一些展望

1.1 什么是计算机视觉

正像其它学科一样，一个大量人员研究了多年的学科，却很难给出一个严格的定义， 模式识别如此，目前火热的人工智能如此，计算机视觉亦如此。与计算机视觉密切相关的概念有视觉感知（visual perception）,视觉认知(visual cognition),图像和视频理解(image and video understanding). 这些概念有一些共性之处，也有本质不同。从广义上说，计算机视觉就是“赋予机器自然视觉能力”的学科。自然视觉能力，就是指生物视觉系统体现的视觉能力。一则生物自然视觉无法严格定义，在加上这种广义视觉定义又“包罗万象”，同时也不太符合 40 多年来计算机视觉的研究状况，所以这种“广义计算机视觉定义”，虽无可挑剔，但也缺乏实质性内容，不过是一种“循环式游戏定义”而已。实际上，计算机视觉本质上就是研究视觉感知问题。视觉感知，根据维科百基（Wikipedia）的定义，是指对“环境表达和理解中，对视觉信息的组织、识别和解释的过程”。根据这种定义，计算机视觉的目标是对环境的表达和理解，核心问题是研究如何对输入的图像信息进行组织，对物体和场景进行识别，进而对图像内容给予解释。

计算机视觉与人工智能有密切联系，但也有本质的不同。人工智能更强调推理和决策，但至少计算机视觉目前还主要停留在图像信息表达和物体识别阶段。“物体识别和场景理解”也涉及从图像特征的推理与决策，但与人工智能的推理和决策有本质区别。应该没有一个严肃的计算机视觉研究人员会认为 AlphaGo, AlphaZero 是计算机视觉， 但都会认为它们是典型的人工智能内容。

简言之，计算机视觉是以图像（视频）为输入，以对环境的表达（representation）和理解为目标，研究图像信息组织、物体和场景识别、进而对事件给予解释的学科。从目前的研究现状看，目前还主要聚焦在图像信息的组织和识别阶段，对事件解释还鲜有涉及，至少还处于非常初级的阶段。

这里需要强调的是，每个人由于背景不同，偏好不同，知识面不同，对同一问题的观点亦会不同，甚至出现大相径庭的局面。上面为笔者对计算机视觉的理解，也许是片面或错误

的。如不少人认为“纹理分析”是计算机视觉的一个重要研究方向，笔者不敢苟同。另外，很多场合，人们把“图像处理”也认为是“计算机视觉”，这也是不恰当的。图像处理是一门独立的学科，图像处理研究图像去噪、图像增强等内容，输入为图像，输出也是图像。计算机视觉利用图像处理技术进行图像预处理，但图像处理本身构不成计算机视觉的核心内容。

这里顺便说一下，目前很多人对“感知”和“认知”不加区分，给读者带来不必要的困惑和误解。在不少场合下，经常会见到有些“视觉专家”把“认知”和“推理与决策”（reasoning and decision）作为平行概念使用，这事实上是不太严谨的。 根据“维基百科”，“认知” 是指通过感觉（senses）、经历 (experience)和思考(thoughts)来获取知识(knowledge)和进行理解(understanding)的思维过程（mental process）。认知包括：知识形成 （knowledge）， 注视（attention），记忆（memory），推理（reasoning），问题求解（problem solving）、决策（ decision making）以及语言生成（ language production）等。所以，“感知”与“认知”有区别，推理和决策是典型的认知过程，是认知的重要组成部分，它们之间是包含关系，不是平行关系。

1.2 计算机视觉发展的四个主要阶段

尽管人们对计算机视觉这门学科的起始时间和发展历史有不同的看法，但应该说，1982 年马尔(David Marr)《视觉》（Marr, 1982）一书的问世，标志着计算机视觉成为了一门独立学科。计算机视觉的研究内容，大体可以分为物体视觉（object vision）和空间视觉（spatial vision）二大部分. 物体视觉在于对物体进行精细分类和鉴别，而空间视觉在于确定物体的位置和形状，为“动作（action）”服务。正像著名的认知心理学家 J.J. Gibson 所言，视觉的主要功能在于“适应外界环境，控制自身运动”。 适应外界环境和控制自身运动，是生物生存的需求，这些功能的实现需要靠物体视觉和空间视觉协调完成。

计算机视觉 40 多年的发展中，尽管人们提出了大量的理论和方法，但总体上说，计算机视觉经历了 4 个主要历程。即： 马尔计算视觉、 主动和目的视觉、多视几何与分层三维重建和基于学习的视觉。下面将对这 4 项主要内容进行简要介绍。

1.2.1 马尔计算视觉（ Computational Vision ）

现在很多计算机视觉的研究人员，恐怕对“马尔计算视觉”根本不了解，这不能不说是一件非常遗憾的事。目前，在计算机上调“深度网络”来提高物体识别的精度似乎就等于从事“视觉研究”。事实上，马尔的计算视觉的提出，不论在理论上还是研究视觉的方法论上，均具有划时代的意义。

马尔的计算视觉分为三个层次： 计算理论、表达和算法以及算法实现。由于马尔认为算法实现并不影响算法的功能和效果，所以，马尔计算视觉理论主要讨论“计算理论”和“表达与算法”二部分内容。马尔认为，大脑的神经计算和计算机的数值计算没有本质区别，所以马尔没有对“算法实现”进行任何探讨。从现在神经科学的进展看，“神经计算”与数值计算在有些情况下会产生本质区别，如目前兴起的神经形态计算（ Neuromorphological computing），但总体上说，“数值计算”可以“模拟神经计算”。至少从现在看，“算法的不同实现途径”，并不影响马尔计算视觉理论的本质属性。

计算理论 (Computational Theory)

计算理论需要明确视觉目的， 或视觉的主要功能是什么。上世纪 70 年代，人们对大脑的认识还非常粗浅，目前普遍使用的非创伤型成像手段，如功能核磁共振（FMRI）等，还没有普及。所以，人们主要靠病理学和心理学结果来推断生理功能。即使目前，人们对“视觉的主要功能”到底是什么，也仍然没有定论。如最近几年，MIT 的 DiCarlo 等人提出了所谓的“目标驱动的感知信息建模”方法（Yamins &DiCarlo et al. 2016a）。他们猜测，猴子 IT 区（IT: interior temporal cortex, 物体识别区）的神经元对物体的响应（neuronal responses）“可以通过层次化的卷积神经网络”（HCNN: Hierarchical Convolutional Neural Networks ）来建模。他们认为，只要对 HCNN 在图像物体分类任务下进行训练，则训练好的 HCNN 可以很好定量预测 IT 区神经元的响应（Yamins et al. 2014, 2016b）。由于仅仅“控制图像分类性能”对 IT 神经元响应（群体神经元对某一输入图像物体的响应，就是神经元对该物体的表达或编码）进行定量预测，所以他们将这种框架称之为“目标驱动的框架”。目标驱动的框架提供了一种新的比较通用的建模群体神经元编码的途径，但也存在很大的不足。能否真正像作者所言的那样，仅仅靠“训练图像分类的 HCNN”就可以定量预测神经元对图像物体的响应，仍是一个有待进一步深入研究的课题。

马尔认为视觉不管有多少功能，主要功能在于“从视网膜成像的二维图像来恢复空间物体的可见三维表面形状”，称之为“三维重建” （3D reconstruction）。而且，马尔认为，这种重建过程不是天生就有的，而是可以通过计算完成的。J.J. Gibson 等心理学家，包括格式塔心里学学派(Gestalt psychology)，认为视觉的很多功能是天生就有的。可以想想，如果一种视觉功能与生具有，不可建模，就谈不上计算，也许就不存在今天的“计算机视觉”这门学科了。

那么，马尔的计算理论是什么呢？这一方面，马尔在其书中似乎并不是介绍得特别具体。

他举了一个购买商品的例子，说明计算理论的重要性。如商店结账要用加法而不是乘法。试想如果用乘法结账，每个商品 1 元钱，则不管你购买多少件商品，你仅仅需要付一元钱。

马尔的计算理论认为，图像是物理空间在视网膜上的投影，所以图像信息蕴含了物理空间的内在信息，因此，任何计算视觉计算理论和方法都应该从图像出发，充分挖掘图像所蕴含的对应物理空间的内在属性。也就是说，马尔的视觉计算理论就是要“挖掘关于成像物理场景的内在属性来完成相应的视觉问题计算”。因为从数学的观点看，仅仅从图像出发，很多视觉问题具有“歧义性”，如典型的左右眼图像之间的对应问题。如果没有任何先验知识，图像点对应关系不能唯一确定。不管任何动物或人，生活的环境都不是随机的，不管有意识或无意识，时时刻刻都在利用这些先验知识，来解释看到的场景和指导日常的行为和行动。如桌子上放一个水杯的场景，人们会正确地解释为桌子上放了一个水杯，而不把他们看作一个新物体。当然，人类也会经常出错，如大量错觉现象。从这个意义上来说，让计算机来模仿人类视觉是否一定是一条好的途径也是一个未知的命题。飞机的飞行需要借助空气动力学知识，而不是机械地模仿鸟如何飞。

表达和算法 （Representation and Algorithm）

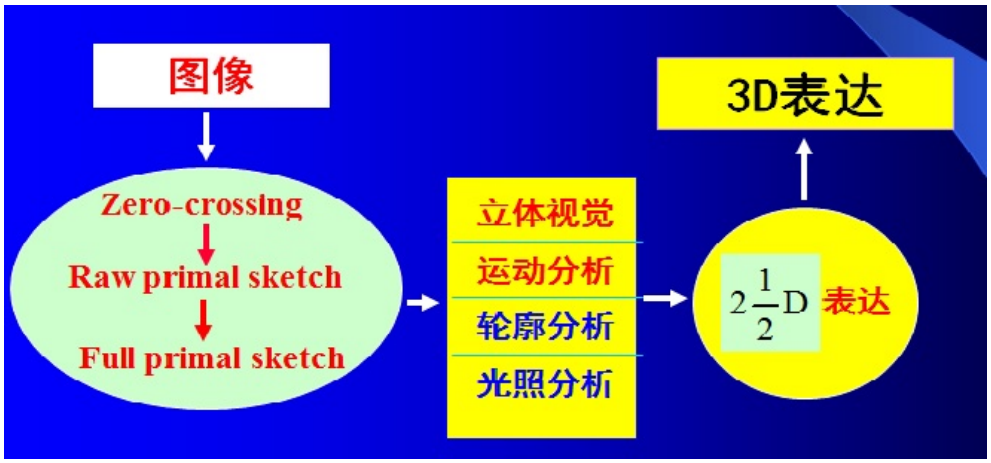
识别物体之前，不管是计算机还是人，大脑（或计算机内存）中事先要有对该物体的存储形式，称之为物体表达（object representation）. 马尔视觉计算理论认为，物体的表达形式为该**物体的三维几何形状**。马尔当时猜测，由于人在识别物体时与观察物体的视角无关，而不同视角下同一物体在视网膜上的成像又不同，所以物体在大脑中的表达不可能是二维的，可能是三维形状，因为三维形状不依赖于观察视角。另外，当时病理学研究发现，有些病人无法辨认“茶杯”，但可以毫无困难地画出茶杯的形状，因此马尔觉得，这些病人也佐证了他的猜测。从目前对大脑的研究看，大脑的功能是分区的。物体的“几何形状”和“语义”储存在不同的脑区。另外，物体识别也不是绝对地与视角无关，仅仅在一个比较小的变化范围内与视角无关。所以，从当前的研究看，马尔的物体的“三维表达”猜测基本上是不正确的，至少是不完全正确的，但马尔的计算理论仍具有重要的理论意义和应用价值。

简言之，马尔视觉计算理论的“物体表达”，是指“**物体坐标系下的三维形状表达**”。注意，从数学上来说，一个三维几何形状，选取的坐标系不同，表达函数亦不同。如一个球体，如果以球心为坐标原点，则球面可以简单表达为 $x^2 + y^2 + z^2 = 1$。 但如果观测者在 x 轴上 2 倍半径处观测，则可见球面部分在观测者坐标系下的方程为：$x = 2 - \sqrt{1 - y^2 - z^2}$. 由此可见，同一物体，选用的坐标系不同，表达方式亦不同。马尔将“观测者坐标系下的三维

几何形状表达”称之为“2.5 维表达”，物体坐标系下的表达为“三维表达”。所以，在后续的算法部分，马尔重点研究了如何从图像先计算“2.5 维表达”，然后转化为“三维表达”的计算方法和过程。

算法部分是马尔计算视觉的主体内容。马尔认为，从图像到三维表达，要经过三个计算层次：首先从图像得到一些基元（primal sketch），然后通过立体视觉（stereopsis）等模块将基元提升到 2.5 维表达，最后提升到 三维表达。

下图总结给出了马尔视觉计算理论的算法流程：



马尔计算理论中算法的三个计算层次

由上图所示，首先从图像提取边缘信息（二阶导数的过零点），然后提取点状基元（blob, 线状基元（edge）和杆状基元 (bar), 进而对这些初级基元（raw primal sketch）组合形成完整基元（full primal sketch）， 上述过程为视觉计算理论的特征提取阶段。在此基础上，通过立体视觉和运动视觉等模块，将基元提升到 2.5 维表达。最后，将 2.5 维表达提升到三维表达。在马尔的《视觉》一书中，重点介绍了特征提取和 2.5 维表达对应的计算方法。在 2.5 维表达部分，也仅仅重点介绍了立体视觉和运动视觉部分。由于当双眼（左右相机）的相互位置已知时（计算机视觉中称之为相机外参数），立体视觉就转化为“左右图像点的对应问题”（image point correspondence），所以，马尔在立体视觉部分重点介绍了图像点之间的匹配问题，即如何剔除误匹配，并给出了对应算法。

立体视觉等计算得到的三维空间点仅仅是在“观测者坐标系下的坐标”，是物体的 2.5 维表示。如何进一步提升到物体坐标系下的三维表示，马尔给出了一些思路，但这方面都很粗泛。如确定物体的旋转主轴等等，这部分内容，类似于后来人们提出的“骨架模型”（skeleton model）构造.

需要指出的是，马尔的视觉计算理论是一种理论体系。在此体系下，可以进一步丰富具体的计算模块，构建“通用性视觉系统”（general vision system）。只可惜马尔（Jan.15,1945 ~ Nov.17,1980 ）1980 年底就因白血病去世，包括他的《视觉》一书，也是他去世后出版的。马尔的英年早逝，不能说不是计算机视觉界的一大损失。由于马尔的贡献，所以二年一度的国际计算机视觉大会（ICCV: International Conference on Computer Vision）设有马尔奖（Marr Prize），作为会议的最佳论文奖。另外，在认知科学领域，也设有马尔奖，因为马尔对认知科学也有巨大的贡献。以同一人名在不同领域设立奖项，实属罕见，可见马尔对计算机视觉的影响有多深远。正如 S. Edelman 和 L. M. Vaina 在 《 International Encyclopedia of the Social & Behavioral Sciences 》中对马尔的评价那样，“马尔前期给出的集成数学和神经生物学对大脑理解的三项工作，已足以使他在任何情况下在英国经验主义二个半世纪的科学殿堂中占有重要的一席，..., 然而，他进一步提出了更加有影响的计算视觉理论”。所以，从事计算机视觉研究的人员对马尔计算视觉不了解，实在是一件比较遗憾的事。

1.2.2 昙花一现的主动和目的视觉

很多人介绍计算机视觉时，将这部分内容不作为一个单独部分加以介绍，主要是因为“主动视觉和目的视觉”并没有对计算机视觉后续研究形成持续影响。但作为计算机视觉发展的一个重要阶段，这里还是有必要予以介绍一下。

上世纪 80 年代初马尔视觉计算理论提出后，学术界兴起了“计算机视觉”的热潮。人们想到的这种理论的一种直接应用就是给工业机器人赋予视觉能力，典型的系统就是所谓的“基于部件的系统”（parts-based system）。然而，10 多年的研究，使人们认识到，尽管马尔计算视觉理论非常优美，但“鲁棒性”（Robustness）不够，很难想人们预想的那样在工业界得到广泛应用。这样，人们开始质疑这种理论的合理性，甚至提出了尖锐的批评。

对马尔计算视觉理论提出批评最多的有二点：一是认为这种三维重建过程是“纯粹自底向上的过程”（pure bottom-up process），缺乏高层反馈（top-down feedback）；二是“重建”缺乏“目的性和主动性”。由于不同的用途，要求重建的精度不同，而不考虑具体任务，仅仅“盲目地重建一个适合任何任务的三维模型”似乎不合理。

对马尔视觉计算理论提出批评的代表性人物有：马里兰大学的 J. Y. Aloimonos; 宾夕法尼亚大学的 R. Bajcsy 和密西根州立大学的 A. K. Jaini。 Bajcsy 认为，视觉过程必然存在人与环境的交互，提出了主动视觉的概念（active vision）. Aloimonos 认为视觉要有目的性，且在很多应用，不需要严格三维重建，提出了“目的和定性视觉”（purpose and qualitative vision）

的概念。 Jain 认为应该重点强调应用，提出了“应用视觉”（ practicing vision）的概念。上世纪 80 年代末到 90 年代初，可以说是计算机视觉领域的“彷徨”阶段。真有点“批评之声不绝，视觉之路茫茫”之势。

针对这种情况，当时视觉领域的一个著名刊物（CVGIP: Image Understanding）于 1994 年组织了一期专刊对计算视觉理论进行了辩论。首先由耶鲁大学的 M. J. Tarr 和布朗大学的 M. J.Black 写了一篇非常有争议性的观点文章（Tarr & Black, 1994），认为马尔的计算视觉并不排斥主动性，但把马尔的“通用视觉理论”（general vision）过分地强调“应用视觉”是“短见”（myopic）之举。通用视觉尽管无法给出严格定义，但“人类视觉”是最好的样板。这篇观点文章发表后，国际上 20 多位著名的视觉专家也发表了他们的观点和评论。大家普遍的观点是，“主动性”“目的性”是合理的，但问题是如何给出新的理论和方法。而当时提出的一些主动视觉方法，一则仅仅是算法层次上的改进，缺乏理论框架上的创新，另外，这些内容也完全可以纳入到马尔计算视觉框架下。所以，从 1994 年这场视觉大辩论后，主动视觉在计算机视觉界基本没有太多实质性进展。这段“彷徨阶段”持续不长，对后续计算机视觉的发展产生的影响不大，犹如“昙花一现”之状。

值得指出的是，“主动视觉”应该是一个非常好的概念，但困难在于“如何计算”。 主动视觉往往需要“视觉注视”（visual attention），需要研究脑皮层（cerebral cortex）高层区域到低层区域的反馈机制，这些问题，即使脑科学和神经科学已经较 20 年前取得了巨大进展的今天，仍缺乏“计算层次上的进展”可为计算机视觉研究人员提供实质性的参考和借鉴。近年来，各种脑成像手段的发展，特别是 “连接组学”（Connectomics）的进展，可望为计算机视觉人员研究大脑反馈机制提供“反馈途径和连接强度”提供一些借鉴。

1.2.3 多视几何和分层三维重建 （multiple View Geometry and Stratified 3D Reconstruction）

上世纪 90 年代初计算机视觉从“萧条”走向进一步“繁荣”，主要得益于以下二方面的因素：首先，瞄准的应用领域从精度和鲁棒性要求太高的“工业应用”转到要求不太高，特别是仅仅需要“视觉效果”的应用领域，如远程视频会议（teleconference），考古，虚拟现实，视频监控等。另一方面，人们发现，多视几何理论下的分层三维重建能有效提高三维重建的鲁棒性和精度。

多视几何的代表性人物首数法国 INRIA 的 O. Faugeras (Faugeras O, 1993), 美国 GE 研究院的 R. Hartely （ 现已回到了澳大利亚国立大学） 和英国牛津大学的 A. Zisserman。应该说，多视几何的理论于 2000 年已基本完善。2000 年 Hartley 和 Zisserman 合著的书 (Hartley

& Zisserman 2000) 对这方面的内容给出了比较系统的总结，而后这方面的工作主要集中在如何提高“大数据下鲁棒性重建的计算效率”。大数据需要全自动重建，而全自动重建需要反复优化，而反复优化需要花费大量计算资源。所以，如何在保证鲁棒性的前提下快速进行大场景的三维重建是后期研究的重点。举一个简单例子，假如要三维重建北京中关村地区，为了保证重建的完整性，需要获取大量的地面和无人机图像。假如获取了 5 百万幅地面高分辨率图像（4000*3000），5 万幅高分辨率无人机图像（8000*7000）（这样的图像规模是当前的典型规模），三维重建要匹配这些图像，从中选取合适的图像集，然后对相机位置信息进行标定并重建出场景的三维结构，如此大的数据量，人工干预是不可能的，所以整个三维重建流程必须全自动进行。这样需要重建算法和系统具有非常高的鲁棒性，否则根本无法全自动三维重建。在鲁棒性保证的情况下，三维重建效率也是一个巨大的挑战。所以，目前在这方面的研究重点是如何快速、鲁棒地重建大场景。

多视几何 （ Multiple View Geometry）

由于图像的成像过程是一个中心投影过程（perspective projection），所以“多视几何”本质上就是研究射影变换下图像对应点之间以及空间点与其投影的图像点之间的约束理论和计算方法的学科（注意：针孔成像模型（The pinhole camera model）是一种中心投影， 当相机有畸变时，需要将畸变后的图像点先校正到无畸变后才可以使用多视几何理论）。计算机视觉领域，多视几何主要研究二幅图像对应点之间的对极几何约束（epipolar geometry），三幅图像对应点之间的三焦张量约束（tri-focal tensor），空间平面点到图像点，或空间点为平面点投影的多幅图像点之间的单应约束（homography）等。在多视几何中，射影变换下的不变量，如绝对二次曲线的像（The image of the absolute conic）,绝对二次曲面的像（The image of the absolute quadric），无穷远平面的单应矩阵（infinite homography），是非常重要的概念，是摄像机能够自标定的“参照物”。由于这些量是无穷远处“参照物”在图像上的投影，所以这些量与相机的位置和运动无关（原则上任何有限的运动不会影响无限远处的物体的性质），所以可以用这些“射影不变量”来自标定摄像机。关于多视几何和摄像机自标定的详细内容，可参阅 Hartley 和 Zisserman 合著的书（Hartley & Zisserman,2000）.

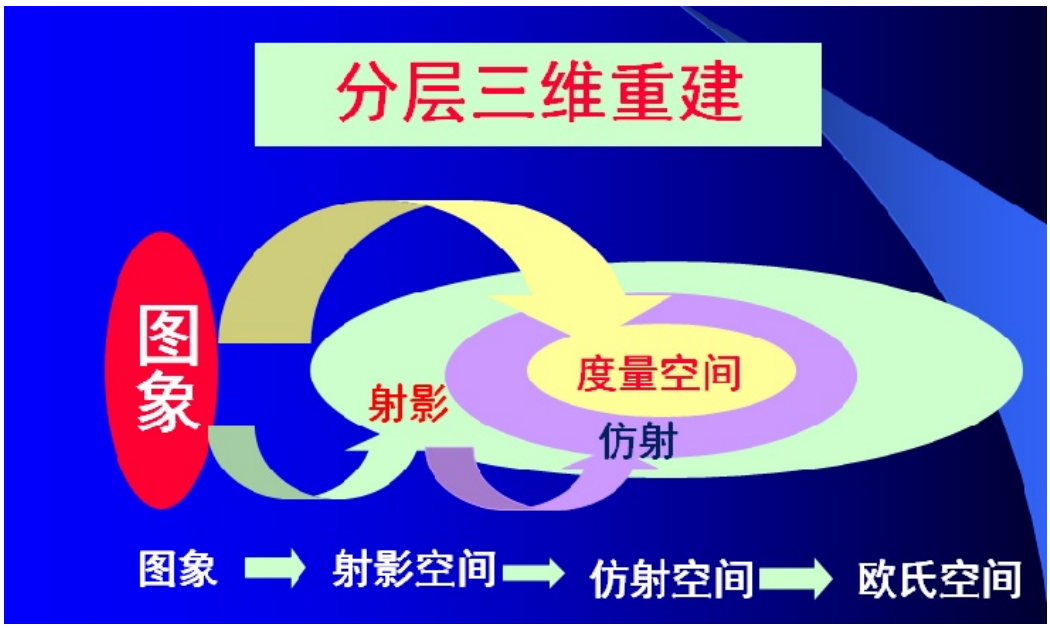
总体上说，多视几何就其理论而言，在射影几何中不能算新内容。Hartley, Faugeras, Zissermann 等将多视几何理论引入到计算机视觉中，提出了分层三维重建理论和摄像机自标定理论，丰富了马尔三维重建理论，提高了三维重建的鲁棒性和对大数据的适应性，有力推动了三维重建的应用范围。所以，计算机视觉中的多视几何研究，是计算机视觉发展

历程中的一个重要阶段和事件。

多视几何需要射影几何（projective geometry）的数学基础。射影几何是非欧几何，涉及平行直线相交，平行平面相交等抽象概念，表达和计算要在“齐次坐标”（homogeneous coordinates）下进行，这给“工科学生”带来不小的困难。所以，大家要从事这方面的研究，一定要先打好基础，至少要具备必要的射影几何知识。否则，做这方面的工作，无异于浪费时间。

分层三维重建 （ Stratified 3D Reconstruction）

所谓的分层三维重建，如下图所示，就是指从多幅二维图像恢复欧几里德空间的三维结构时，不是从图像一步到欧几里德空间下的三维结构，而是分步分层地进行。即先从多幅图像的对应点重建射影空间下的对应空间点(即射影重建：projective reconstruction)，然后把射影空间下重建的点提升到仿射空间下(即仿射重建：affine reconstruction)，最后把仿射空间下重建的点再提升到欧几里德空间（或度量空间: metric reconstruction）（注：度量空间与欧几里德空间差一个常数因子。由于分层三维重建仅仅靠图像进行空间点重建，没有已知的“绝对尺度”，如“窗户的长为 1 米”等，所以从图像仅仅能够把空间点恢复到度量空间）。



这里有几个概念需要解释一下。以空间三维点的三维重建为例，所谓的“射影重建”，是指重建的点的坐标与该点在欧几里德空间下的坐标差一个“射影变换”。所谓的“仿射重

建”，是指重建的点的坐标与该点在欧几里德空间下的坐标差一个“仿射变换”。所谓的“度量重建”，是指重建的点的坐标与该点在欧几里德空间下的坐标差一个“相似变换”。

由于任何一个视觉问题最终都可以转化为一个多参数下的非线性优化问题，而非线性优化的困难在于找到一个合理的初值。由于待优化的参数越多，一般来说解空间越复杂，寻找合适的初值越困难，所以，如果一个优化问题如能将参数分组分步优化，则一般可以大大简化优化问题的难度。分层三维重建计算上的合理性正是利用了这种“分组分步”的优化策略。以三幅图像为例，直接从图像对应点重建度量空间的三维点需要非线性优化 16 个参数（假定相机内参数不变，5 个相机内参数，第二幅和第三幅图像相对于第一幅图像的相机的旋转和平移参数，去掉一个常数因子，所以：５＋２×（３＋３）－１　＝１６），　这是一个非常困难的优化问题。但从图像对应点到射影重建需要“线性”估计２２个参数，由于是线性优化，所以优化问题并不困难。从射影重建提升到仿射重建需要“非线性”优化三个参数（无穷远平面的 3 个平面参数），而从仿射重建提升到度量重建需要“非线性”优化 5 个参数（摄像机的 5 个内参数）。因此，分层三维重建仅仅需要分步优化 3 个和 5 个参数的非线性优化问题，从而大大减小了三维重建的计算复杂度。

分层三维重建的另一个特点是其理论的优美性。射影重建下，空间直线的投影仍为直线，二条相交直线其投影直线仍相交，但空间直线之间的平行性和垂直性不再保持。仿射重建下可以保持直线的平行性，但不能保持直线的垂直性。度量重建既可以保持直线之间的平行线，也可以保持垂直性。在具体应用中，可以利用这些性质逐级提升重建结果。

分层三维重建理论可以说是计算机视觉界继马尔计算视觉理论提出后又一个最重要和最具有影响力的理论。目前很多大公司的三维视觉应用，如苹果公司的三维地图，百度公司的三维地图，　诺基亚的 Streetview, 微软的虚拟地球，其后台核心支撑技术的一项重要技术就是分层三维重建技术。

摄像机自标定　（camera self-calibration）

所谓摄像机标定，狭义上讲，就是确定摄像机内部机械和光电参数的过程，如焦距，光轴与像平面的交点等。尽管相机出厂时都标有一些标准参数，但这些参数一般不够精确，很难直接在三维重建和视觉测量中应用。所以，为了提高三维重建的精度，需要对这些相机内参数（intrinsic parameters）进行估计。估计相机的内参数的过程，称为相机标定。在文献中，有时把估计相机在给定物体坐标系下的坐标，或相机之间相互之间的位置关系，称为相机外参数（extrinsic parameters）标定。但一般无明确指定时，相机标定就是指对相机内参

数的标定。

相机标定包含二方面的内容：“成像模型选择”和“模型参数估计”。相机标定时首先需要确定“合理的相机成像模型”，如是不是针孔模型，有没有畸变等。目前关于相机模型选择方面，没有太好的指导理论，只能根据具体相机和具体应用确定。随着相机加工工艺的提高，一般来说，普通相机（非鱼眼或大广角镜头等特殊相机）一般使用针孔成像模型（ 加一阶或二阶径向畸变）就足以了。其它畸变很小，可以不加考虑。当相机成像模型确定后，进一步需要估计对应的模型参数。文献中人们往往将成像模型参数估计简单地认为就是相机标定，是不全面的。事实上，相机模型选择是相机标定最关键的步骤。一种相机如果无畸变而在标定时考虑了畸变，或有畸变而未加考虑，都会产生大的误差。视觉应用人员应该特别关注“相机模型选择”问题。

相机参数估计原则上均需要一个“已知三维结构”的“标定参考物”，如平面棋盘格，立体块等。所谓相机标定，就是利用已知标定参考物和其投影图像，在已知成像模型下建立模型参数的约束方程，进而估计模型参数的过程。所谓“自标定”，就是指“仅仅利用图像特征点之间的对应关系，不需要借助具体物理标定参考物，进行模型参数估计的过程”。“传统标定”需要使用加工尺寸已知的标定参考物，自标定不需要这类物理标定物，正像前面多视几何部分所言，使用的是抽象的无穷远平面上的“绝对二次曲线”和“绝对二次曲面”。从这个意义上来说，自标定也需要参考物，仅仅是“虚拟的无穷远处的参考物”而已。

摄像机自标定需要用到两幅图像之间的约束，如基础矩阵（fundamental matrix），本质矩阵（essential matrix），以及三幅图像之间的三焦张量约束等。另外，Kruppa 方程也是一个重要的概念。这些内容是多视几何的重要内容，后续章节将进行详细介绍。

1.2.4 基于学习的视觉（ learning based vision）

基于学习的视觉，是指以机器学习为主要技术手段的计算机视觉研究。基于学习的视觉研究，文献中大体上分为二个阶段：本世纪初的以流形学习(manifold Learning)为代表的子空间法(subspace method)和目前以深度神经网络和深度学习（deep neural networks and deep learning）为代表的视觉方法。

流形学习 （Manifold Learning）

正像前面所指出的，物体表达是物体识别的核心问题。给定图像物体，如人脸图像，不同的表达，物体的分类和识别率不同。另外，直接将图像像素作为表达是一种“过表达”，

也不是一种好的表达。流形学习理论认为，一种图像物体存在其“内在流形”（intrinsic manifold），这种内在流形是该物体的一种优质表达。所以，流形学习就是从图像表达学习其内在流形表达的过程，这种内在流形的学习过程一般是一种非线性优化过程。

流形学习始于 2000 年在 Science 上发表的二篇文章（ Tenenbaum et al., 2000） (Roweis & Lawrence 2000)。流形学习一个困难的问题是没有严格的理论来确定内在流形的维度。人们发现，很多情况下流形学习的结果还不如传统的 PCA （Principal Component Analysis），LDA （ linear Discriminant Analysis ）， MDS（ Multidimensional Scaling）等. 流形学习的代表方法有：LLE （Locally Linear Embedding ）(Roweis & Lawrence 2000)，Isomap （ Tenenbaum et al., 2000）， Laplacian Eigenmaps (Belkin & Niyogi, 2001) 等。

深度学习（Deep Learning ）

深度学习(LeCun et al. 2015) 的成功，主要得益于数据积累和计算能力的提高。深度网络的概念上世纪 80 年代就已提出来了，只是因为当时发现“深度网络”性能还不如“浅层网络”，所以没有得到大的发展。目前似乎有点计算机视觉就是深度学习的应用之势，这可以从计算机视觉的三大国际会议：国际计算机视觉会议（ICCV），欧洲计算机视觉会议（ECCV）和计算机视觉和模式识别会议（CVPR）， 上近年来发表的论文可见一般。目前的基本状况是，人们都在利用深度学习来“取代”计算机视觉中的传统方法。“研究人员”成了“调程序的机器”，这实在是一种不正常的“群众式运动”。牛顿的万有引力定律，麦克斯韦的电磁方程，爱因斯坦的质能方程，量子力学中的薛定谔方程，似乎还是人们应该追求的目标。

关于深度网络和深度学习，详细内容可参阅相关文献，这里仅仅强调以下几点：

（1）深度学习在物体视觉方面较传统方法体现了巨大优势，但在空间视觉，如三维重建，物体定位方面，仍无法与基于几何的方法相媲美。这主要是因为深度学习很难处理图像特征之间的误匹配现象。在基于几何的三维重建中，RANSAC （Random Sample Consensus）等鲁棒外点（误匹配点）剔除模块可以反复调用，而在深度学习中，目前还很难集成诸如 RANSAC 等外点剔除机制。笔者认为，如果深度网络不能很好地集成外点剔除模块，深度学习在三维重建中将很难与基于几何的方法相媲美，甚至很难在空间视觉中得到有效应用；

（2） 深度学习在静态图像物体识别方面已经成熟，这也是为什么在 ImageNet 上的物体分类竞赛已不再举行的缘故；

（3） 目前的深度网络，基本上是前馈网络（feedforward Networks）.不同网络主要体

现在使用的代价函数不同。下一步预计要探索具有“反馈机制”的层次化网络。反馈机制，需要借鉴脑神经网络机制，特别是连接组学的成果。

（4） 目前对视频的处理，人们提出了 RCNN (Recurrent Neural Networks). 循环 (recurrent) 是一种有效的同层作用机制，但不能代替反馈。大脑皮层远距离的反馈（将在生物视觉简介一章介绍)可能是形成大脑皮层不同区域具有不同特定功能的神经基础。所以，研究反馈机制，特别具有“长距离反馈”（跨多层之间）的深度网络，将是今后研究图像理解的一个重要方向；

（5）尽管深度学习和深度网络在图像物体识别方面取得了“变革性”成果，但为什么“深度学习”会取得如此好的结果目前仍然缺乏坚实的理论基础。目前已有一些这方面的研究，但仍缺乏系统性的理论。事实上，“层次化”是本质，不仅深度网络，其它层次化模型，如 Hmax 模型（Riesenhuber & Poggio,1999） HTM （ Hierarchical Temporal memory）模型（George & Hawkins, 2009）存在同样的理论困惑。为什么“层次化结构”（ hierarchical structure ）具有优势仍是一个巨大的迷。

1.3 计算机视觉的若干发展趋势

信息科学发展之迅速，对未来 10 年的发展趋势进行预测，有点“算命”的感觉。 对计算机视觉而言，笔者有以下几点对未来发展的展望：

（1） 基于学习的物体视觉和基于几何的空间视觉继续“相互独立”进行。深度学习在短时期内很难代替几何视觉。在深度网络中如何引入“鲁棒外点剔除模块”将是一个探索方向，但短时间内估计很难有实质性进展；

（2） 基于视觉的定位将更加趋向“应用性研究”，特别是多传感器融合的视觉定位技术。

（3） 三维点云重建技术已经比较成熟，如何从“点云”到“语义”是未来研究重点。“语义重建”将点云重建、物体分割和物体识别同时进行，是三维重建走向实用的前提。

（4）对室外场景的三维重建，如何重建符合“城市管理规范”的模型是一个有待解决的问题。室内场景重建估计最大的潜在应用是“家庭服务机器人”。 鉴于室内重建的应用还缺乏非常具体的应用需求和驱动，在加上室内环境的复杂性，估计在 3 — 5 年内很难有突破性进展。

（5）对物体识别而言，基于深度学习的物体识别估计将从“通用识别”向“特定领域物体的识别”发展。“特定领域”可以提供更加明确和具体的先验信息，可以有效提高识别的精度和效率，更加具有实用性；

（6）目前基于 RCNN 对视频理解的趋势将会持续；

（7）解析深度网络机理的工作具有重大的理论意义和挑战性，鉴于深度网络的复杂性，估计近期很难取得突破性进展；

（8）具有“反馈机制”的深度网络结构（architecture）研究必将是下一个研究热点。

1.4 几种典型的物体表达理论 （ Some object representation theories）

正像前面所述，物体表达是计算机视觉的一个核心科学问题。这里，“物体表达理论”与“物体表达模型”需要加以区别。“表达理论”是指文献中大家比较认可的方法。“表达模型”容易误解为“数学上对物体的某种描述”。计算机视觉领域，比较著名的物体表达理论有以下三种：

（1） 马尔的三维物体表达

前面已经介绍过，马尔视觉计算理论认为物体的表达是物体坐标系下的三维表达

（2） 基于二维图像的物体表达（View-based object representation）

尽管理论上一个三维物体可以成像为无限多不同的二维图像，但人的视觉系统仅仅可以识别“有限个图像”。鉴于神经科学对于猴子腹部通道（ventral pathway）(注：腹部通道认为是物体识别通道)的研究进展，T. Poggio 等提出了基于图像的物体表达（Poggio & Bizzi, 2004），即对一个三维物体的表达是该物体的一组典型的二维图像（view）。目前，也有人认为 Poggio 等的 ”view”不能狭义地理解为二维图像，也包含以观测者为坐标系下的三维表示，即马尔的 2.5 维表示（Anzai & DeAngelis，2010）。

（3）逆生成模型表达（ inverse generative model representation ）

长期以来，人们认为物体识别模型为“鉴别模型”（ discriminative model），而不是“生成模型”（ generative model ）。近期对猴子腹部通道的物体识别研究表明，猴子大脑皮层的 IT 区（ Inferior Temporal: 物体表达区域）可能在于编码物体及其成像参数（如光照和姿态，几何形状，纹理等）（Yildirim et al. 2015）（Yamins &DiCarlo，2016b.）。 由于已知这些参数就可以生成对应图像，所以对这些参数的编码可以认为是逆生成模型表达。逆生成模型

表达可以解释为什么深度学习中的 Encoder-decoder 网络结(Badrinarayanan et al. 2015) 可以取得比较好的效果，因为 Encoder 本质上就是图像的逆生成模型。另外，深度学习中提出的“逆图形学”概念（ Inverse Graphic）(Kulkarni et al. 2015), 从原理上也是一种逆生成模型。逆图形学是指先从图像学习到图像生成参数，然后把同一物体在不同参数下的图像归类为同一物体，通过这种 “等变物体识别”(Equivariant recognition) 来达到最终的“不变物体识别”（invariant recognition）。

总之，本文对计算机视觉的理论、现状和未来发展趋势进行了一些总结和展望，希望能给读者了解该领域提供一些帮助。特别需要指出的是，这里很多内容也仅仅是笔者的一些“个人观点”和“个人偏好”下总结的一些内容，以期对读者有所帮助但不引起误导。另外，笔者始终认为，任何一门学科的核心关键文献并不多，为了读者阅读方便，所以本文也仅仅给出了一些必要的代表性文献。

1. Marr D (1982), Vision: A computational investigation into the human representation and processing of visual information, W.H. Freeman and Company.
2. Yamins D. L K. & DiCarlo J.J (2016a)， Using goal-driven deep learning models to understand sensory cortex， Nature Neuroscience，Perspective， Vol. 19， No.3, pp.356-365..
3. Yamins D. L. K et al.(2014), Performance-optimized hierarchical models predict neural responses in higher visual cortex, PNAS, Vol.111, No.23, pp.8619-8624.
4. Yamins D. L. K & DiCarlo J. J. (2016b). Explicit Information for category-orthogonal object properties increases along the ventral stream, Nature Neuroscience, Vo.19, No.4, pp.613-622.
5. Edelman. S & Vaina L. (2015). Marr, David (1945–80),International Encyclopedia of the Social & Behavioral Sciences (Second Edition), pp. 596-598
6. Tarr. M & Black M. (1994). A computational and Evolutionary Perspective on the Role of Representation in Vision, CVGIP: Image Understanding, Vol.60, No.1, pp.65-73.
7. Hartley R & Zisserman A.(2000). Multiple View Geometry in Computer Vision, Cambridge University Press.
8. Faugeras O (1993). Three-Dimensional Computer Vision: A geometric Viewpoint, MIT Press.
9. Tenenbaum J. B. et al. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science, Vol. 290, No.5500, pp.2319–2323.
10. Roweis.S &Saul. L(2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science, Vol.290, No.5500, pp.2323—2326.
11. Belkin.M & Niyogi.P.(2001). Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, Advances in Neural Information Processing Systems 14, pp. 586–691, MIT Press

12. LeCun .Y et al. (2015). Deep Learning, Nature, Vol.521 , pp.436-444.
13. Riesenhuber .M. & Poggio. T.(1999). Hierarchical Models of Object Recognition in Cortex, Nature Neuroscience 2: 1019-1025.
14. George D & Hawkins J (2009). Towards a mathematical theory of cortical micro-circuits, PloS Computational Biology, Vol.5, No.10, pp.1-26.
15. Poggio. T & Bizzi. E(2004). Generalization in vision and motor control, Nature 431, Vol.14, pp.768-774.
16. Anzai. A & DeAngelis. G (2010). Neural computations underlying depth perception，Curr Opin Neurobiol., Vol.20, No.3, pp. 367–375.
17. Yildirim. I et al. (2015). Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations. In Proceedings of the 37th Annual Cognitive Science Society.
18. Badrinarayanan. V et al. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation, arXiv:1511.00561.
19. Kulkarni T. D. et al.(2015). Deep Convolutional Inverse Graphics Network, NIPS 2015