Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

观点：

论视觉知识

潘云鹤

浙江大学，中国杭州市，310027

E-mail: panyh@cae.cn

摘要：提出“视觉知识”概念. 视觉知识是知识表达的一种新形式. 它与迄今为止人工智能（AI）所用知识表达方法不同. 其中视觉概念具有典型（prototype）与范畴结构、层次结构与动作结构等要素. 视觉概念能构成视觉命题，包括场景结构与动态结构，视觉命题能构成视觉叙事. 指出重构计算机图形学成果可实现视觉知识表达及其推理与操作，重构计算机视觉成果可实现视觉知识学习. 实现视觉知识表达、推理、学习和应用技术将是 AI 2.0 取得突破的重要方向之一.

本文译自 Pan YH, 2019. On visual knowledge. *Front Inform Technol Electron Eng*, 20(8):1021-1025.
https://doi.org/10.1631/FITEE.1910001

1 视觉知识对 AI 发展有重要影响

1.1 图像识别水平快速提升引发 AI 热潮

近年来，图像识别水平的快速提升推动 AI 热潮形成. 例如，2012 年，卷积神经网络 AlexNet 在面向 ImageNet 的大规模视觉识别竞赛中以明显优势（相比于第 2 名 26.2%的错误率，AlexNet 的错误率为 15.3%）战胜了传统机器学习方法，使得深度学习成为学术界和产业界的焦点. 2016 年 5 月美国白宫发表文章《为人工智能的未来做准备》谈到：鉴于人工智能在医学以及图像语音理解等方面将对社会生活起到史无前例的影响，在美国国家科技委中设立“人工智能和机器学习委员会”，协调指导 AI 技术在工业、研究委员会以及联邦政府中的发展.

图像识别技术的突破不仅提高了计算机对人脸、文字、指纹及生物特征、医学图片等识别的准确率，而且进一步推动了智能汽车、安全监控、智能交通、机器人、无人机、智能制造等广泛领域的发展. 图 1 是中国科学技术发展战略研究院对 2018 年中美 AI 企业数量按技术分类的统计. 从中可以看出，提供和应用图像识别技术的企业占一半以上.

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2019

1.2 基于多层神经网络的深度学习模型可看作一种新的知识表达

传统图像识别建立在图像处理技术之上，而图像处理技术可追溯至 19 世纪 20 年代纽约至伦敦的电缆传输新闻数字图像. 1977 年，冈萨雷斯在《数字图像处理》教材中，对图像编辑、变换、增强、分割、边界检测与目标提取进行了系统整理，在此基础上，发展了经典的图像识别与计算机视觉技术.

卷积神经网络（CNN）的使用改变了传统计算机视觉处理方法，以数据驱动方法来学习特征表达，有效提高了图像分类和识别的准确率. CNN 的成功，从 AI 角度可以看作一种新知识表达的成功. 此前，AI 通用知识表达皆为符号型知识表达，包括规则、框架和语义网络等等. 而深度学习获得的则是具有权重的网络性知识表达，如 CNN 网络结构及连接权重.

卷积神经网络具有两大优点：（1）可从标识的样本数据自动学习；（2）可用于非符号数据识别，如图像与语音识别.

但是卷积神经网络也有缺点：（1）不可解释；（2）不可推理；（3）需大量标识的数据训练网络参数，这种大量标识的数据不可避免地会引入数据偏见（Hutchinson and Mitchell, 2019）. 这从另

一角度说明此类知识表达具有显著缺陷. 因此需要研究一种全新的知识表达——视觉知识.

上述分析给予我们的启发是：（1）数字视觉领域是推动 AI 发展的重要领域；（2）更好的知识表达是推动数字视觉发展的关键技术；（3）克服深度神经网络缺陷是视觉知识研究的关键方向.

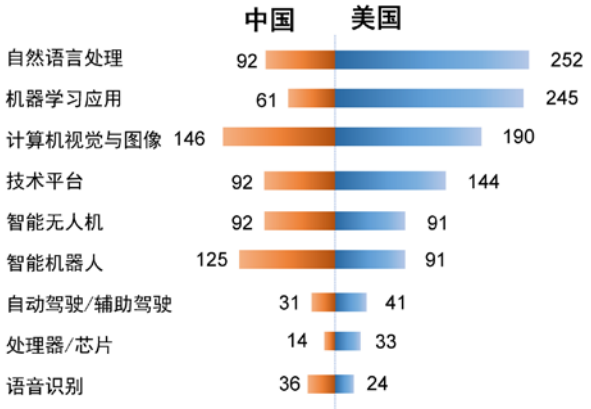

图 1 2018 年中美 AI 企业数量（按技术分类统计）

2 视觉知识具有不可替代性

2.1 认知科学对视觉认知的重要性和独特性做过大量重要研究

上世纪 70 年代，认知心理学家对视觉记忆做过一系列实验，证明视觉记忆与言语记忆不同，它们是可以被旋转、折叠、扫描和类比的. 以下描述了这些实验（Anderson, 1989）.

1. 谢泼德（R. N. Shepard）1971 年心理旋转实验

该实验测试了参与者对三维物体的心理旋转能力. 让参与者依次观看一对三维物体的二维图形，并决定两者是否为同一物体. 实验表明参与者能对自己脑内的三维对象施加旋转与比较，所需反应时间与从原始位置旋转的角度成正比. 也就是说，两个对象之间差异角度越大，参与者完成旋转与比较所需时间越长.

2. 谢泼德等 1972 年心理折纸实验

参与者依次观看一个展开为平面的立方体上的两个箭头（这两个箭头分别被标记在两个不同平面的边缘），并判定将其折叠回立方体时箭头是否相遇. 实验结果表明：参与者能对脑内视觉记忆做折叠，判定所需平均时间是使箭头一致所需的折叠次数的函数.

3. 库斯林（S. M. Kosslyn）等 1978 年心象扫描实验

在该实验中，参与者需将注意力从照片中一个物体转移到另一个物体. 实验结果表明，参与者在两个远距离物体（例如小屋和岩石）之间转移注意力比在两个相邻物体（例如小屋和水井）之间转移注意力所需时间更久. 该心象扫描实验显示在大脑记忆图像中扫描两个物体所需时间与真实图像中两个物体之间的距离存在强线性关联.

4. 莫耶（R. S. Moyer）1973 年记忆中动物大小比较实验

通过在记忆中检索信息，人类能够比较两种物体的差异. 在该实验中，当参与者听到两种动物名字时，他们从记忆中判断这两种动物的相对大小. 实验结果说明，参与者的反应时间随着两者大小差异增加而减少.

上述视觉记忆被认知心理学称为心象. 研究指出，心象是形象思维的知识形式（潘云鹤，1991）. 从 AI 角度，我们将视觉心象称为视觉知识.

2.2 视觉知识的特征

上述认知心理学实验说明人脑记忆中的视觉知识具有一系列特性：（1）能表达对象的空间形状、大小和空间关系，以及色彩和纹理. （2）能表达对象的动作、速度及时间关系. （3）能进行对象的时空变换、操作与推理，包括形状变换、动作变换、速度变换、场景变换，各种时空类比，联想和基于时空推理结果预测，等等.

认知心理学研究还指出：人类记忆的视觉知识远多于言语知识，而言语知识的理解，也不能脱离视觉知识的支持（Horoufchin et al., 2018; Kosiorek et al., 2019）. 视觉知识因为难以用语言符号表达，曾被统归为常识. 以往 AI 研究一大弱点便是视觉知识研究不足. 因此，视觉知识的研究与运用将会是 AI 2.0 的一个重要发展方向（Pan, 2016）.

3 视觉知识的表达与操作：基于计算机图形学的重构

计算机图形学经过长期研究，积累了 3D 形状表达与操作算法，如半边数据结构、几何变换、欧拉操作、投影变换（潘云鹤等，2011），以及动

画和变形等技术. 它们提供了视觉知识表达与操作的技术基础.但是，对视觉知识进行表达及其推理等操作，还需在此基础上加以改造与重构.

3.1 图形表达需重构为视觉概念

3.1.1 视觉概念

视觉概念通常由典型（prototype）和范畴构成. 例如，苹果有千变万化的形状，但必有一种或几种核心形状和色彩，称为典型.围绕核心形状和色彩，各种苹果构成一个变化范围，变化范围会有一个边界.边界内的形状属于苹果范畴，超过这个边界就变成其他水果形状和色彩范畴.所以，视觉概念={典型，范畴}.

范畴既可表达为典型中各种参数的变化域，也可表达为典型和若干非典型形状、色彩所构成的综合场（潘云鹤，1996）.

3.1.2 视觉概念的层次结构

视觉概念应含有子概念空间组织的结构. 如苹果是由果核、果肉、果皮、果蒂等子概念组成的结构. 视觉概念要表达对象的这种空间结构关系.如苹果的这种结构表达在解决植物学、农业和食品方面的问题时，都有用处.

3.1.3 视觉概念的动作结构

动物等视觉概念还应含有动作. 动作应包含结构中各子概念的典型运动及动作范畴. 如动物的头、肢、躯、爪（指）的各自动作及其关系等表达.

3.2 视觉操作与推理

视觉的操作与推理包括形状的分解、替换、组合、变形、运动、比较、综合、破坏、修复、预测等等.它们需要在图形变换、变形、动画等技术基础上重构.

4 视觉知识学习：基于计算机视觉的重构

视觉知识学习是视觉知识体系构建与利用需要解决的首要问题。尽管现有的一些自底向上技术（Greff et al., 2019; Zhao et al., 2019）可以用来获取部分视觉知识，但是目前没有一个系统的视觉知识学习方法.建立一个系统的知识体系，往往更需要自顶向下的设计. 这个过程中计算机视觉研究成果，如 3D 重建，为这种系统的视觉知识学习提供了生长土壤.

4.1 视觉概念学习

计算机视觉经过长期研究，已积累许多技术，包括由多幅图像恢复 3D 物体形状的技术，用 3D 扫描仪扫描真实物体，得到扫描点云，从而创造 3D 形状网格的技术，通过摄像机从不同角度同步拍摄真实物体（及其动作），比较与计算对应点而重建 3D 物体形状的技术（及其动作的动画技术），等等（马颂德等，1998）.

视觉知识学习要将目标在视觉形状重建基础上进一步深入到视觉知识重建，这就需要对现有计算机视觉技术作如下进一步研究：（1）不仅重建 3D 形状，而且重建 3D 形状的层次结构；（2）不仅重建 3D 形状结构，而且定位其在概念范畴中的位置，如典型位置、边缘位置、中间位置等等.

4.2 视觉命题的表达学习

除了视觉概念，还要研究视觉叙事的表达和学习.视觉叙事由一组视觉命题构成，而视觉命题是对视觉概念的空间关系和时间关系的表达.

空间关系表达为场景结构，描述各对象之间上下、左右、前后等方位关系、距离关系、里外关系、大小关系等等.

时间关系表达为动态结构，表达对象的生长、位移、动作、变化、竞赛、协同等等.

一幅图像对应一个描写场景结构的视觉命题. 一段视频则是一个典型的视觉叙事，它会对应一组序贯视觉命题，既有场景结构，又可能有动态结构. 无声电影已证明视觉叙事有很强的表达能力.视觉知识的学习任务，还包含在视觉概念知识基础上自动学习视觉命题知识，以及在视觉命题基础上自动学习视觉叙事知识.

5 视觉知识的使用方法分析

从当前 AI 热潮中视觉识别技术的广泛渗透，可推知视觉知识应用极广.下举 3 例，说明视觉知识应用的各种方法：基于生成的识别方法、基于知识的重建方法、基于知识的生成方法等.

例 1：基于生成的识别——视觉知识用于图像识别（例如猫）

（1）根据“猫”的视觉概念的典型与范畴等，使用综合推理方法（潘云鹤，1996），自动生成猫的范畴内外各种图像大数据，并根据范畴内外自动标识为正、负范例. （2）用上述范例大数据训练多层神经网络. （3）用训练过的多层神经网络识别图像.

例 2：基于知识的重建——视觉知识用于 3D 重建（例如试衣人体）

（1）从试者图像中用机器视觉技术获得人体特征点. （2）在人体概念范畴中取出符合试者参数（如性别、年龄、体重、身高等）的 3D 人体，并以特征点修改之，获得特征人体. （3）将特征人体投影与图像人体比较而获得差异. （4）以差异修改特征人体，返回步骤（3），直至差异可忽略. 可以看出，上述过程是很自然的“知识迁移学习”，可大大减少视觉知识学习工作量.

例 3：基于知识的生成——视觉知识用于设计（例如设计人物角色，在动画、游戏、绘画、广告应用中都有需要）

（1）确定角色要求. 例如，一群跳跃的青年男子，身材与体重各异，肌肉发达，行动敏捷，走跑矫健. （2）根据视觉概念“青年男子”的典型，向范畴的矫健型方向生成各种高度和起跳时间地点的跳跃男子形象. （3）用户对生成的跳跃男子形象提出意见，根据意见返回步骤（2），调整范畴方向再生成跳跃男子形象，直到用户满意.

6 小结

由上述分析可知，视觉知识的独特优点是能够提供综合生成能力、时空比较能力和形象显示能力. 这些正是字符知识所缺乏的重要能力. 它们能在创造、预测和人机融合等方面对 AI 新发展提供新的基础动力. 因此，视觉知识研究将会促进发展新的视觉智能，是促进 AI 2.0 取得重要突破的关键技术之一（Pan, 2016）.

建设视觉知识词典将是十分重要的. 这是一个巨大而实用的知识平台和数据平台，应当联合全球人工智能、计算机图形学和计算机视觉科技工作者共同建设之. 为顺利而高效地完成此建设，群智组织模式也将不可或缺.

致谢

感谢庄越挺、吴飞、汤斯亮教授对本文提出很好建议.

参考文献

Anderson JR, 1989. Cognitive Psychology (Yang Q, Trans.). Jilin Education Press, China.

Greff K, Kaufmann RL, Kabra R, et al., 2019. Multi- object representation learning with iterative variational inference. https://arxiv.org/abs/1903.00450

Horoufchin H, Bzdok D, Buccino G, et al., 2018. Action and object words are differentially anchored in the sensory motor system—a perspective on cognitive embodiment. *Sci Reports*, 8:6583. https://doi.org/10.1038/s41598-018-24475-z

Hutchinson B, Mitchell M, 2019. 50 years of test (un)fairness: lessons for machine learning. Proc Conf on Fairness, Accountability, and Transparency, p.49-58. https://doi.org/10.1145/3287560.3287600

Kosiorek AR, Sabour S, Teh YW, et al., 2019. Stacked capsule autoencoders. https://arxiv.org/abs/1906.06818

马颂德，张正友，1998. 计算机视觉：计算理论与算法基础. 北京：科学出版社.

潘云鹤，1991. 形象思维中的形象信息模型的研究. 模式识别与人工智能，4(4):7-12.

潘云鹤，1996. 综合推理的研究. 模式识别与人工智能，9(3):201-208.

Pan YH, 2016. Heading toward artificial intelligence 2.0. *Engineering*, 2(4):409-413. https://doi.org/10.1016/J.ENG.2016.04.018

潘云鹤，童若锋，唐敏，2011. 计算机图形学——原理、方法及应用（第 3 版）. 北京：高等教育出版社.

Zhao Y, Birdal T, Deng H, et al., 2019. 3D point capsule networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1009-1018.