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：

视觉知识的五个基本问题

潘云鹤

浙江大学计算机学院人工智能研究所，中国杭州市，310027

E-mail: panyh@zju.edu.cn

本文编译自 Pan YH, 2021. Miniaturized five fundamental issues about visual knowledge. *Front Inform Technol Electron Eng*, 22(5):615-618. https://doi.org/10.1631/FITEE.2040000

认知心理学早已指出，视觉知识是人类知识记忆的重要部分，被用来进行形象思维. 因此，基于视觉的人工智能（AI）是 AI 绕不开的课题，且具有重要意义. 本文继《论视觉知识》一文（Pan, 2019），讨论与之相关的 5 个基本问题.

1 基本问题 1：视觉知识表达

认知心理学实验说明了视觉知识不同于言语知识的特征：（1）能表达对象的大小、色彩、纹理、空间形状及关系. （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. 如，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的预测等.

以计算机图形学（CG）方式实现的形体表达较适合表达特征（1）和（3）的操作与变化，而模拟想象变化则较困难，因为 CG 表达的是几何形态，而不是视觉概念.

视觉概念应由典型（prototype）和范畴（domain）构成. 例如，苹果有千变万化的形态，但必有一种或几种核心形状和色彩，称为典型. 围绕典型，各种苹果构成一个变化范围，变化范围会有一个边界. 边界内的形态属于苹果的范畴，超过这个边界就变成其他水果. 这个变化范围就是苹果这一概念的范畴. 视觉概念有层次结构，即含有子概念的空间组织的结构. 视觉概念有动作结构，应包含结构中各子概念的典型运动及动作范畴.

视觉命题是视觉概念的空间关系和时间关系表达. 空间关系表达为场景结构，描述上下、左右、前后等方位关系、距离关系、里外关系、大小关系等几何模式. 时间关系表达为动态结构，表达生长、位移、动作、变化、竞争、协同和演化等时序模式.

视觉叙事由一组视觉命题构成. 如一段视频中不同视觉对象在各种场景中的动态. 无声电影是视觉叙事的例子.

视觉叙事是具象连续表达，言语叙事是抽象离散表达. 哑语是仿语言的视觉表达.

认知心理学研究指出：人类记忆中储存的视觉知识远多于言语知识. 视觉知识因难以用语言符号表达，曾被统归为常识. 如：儿童在 5 岁之前，看到不同杯子就会以不同手法抓来喝水. 证明儿童已能熟练运用视觉知识，但未能用言语解释（图 1）. 人在幼年，学到的多是视觉知识.

潘云鹤院士

以往 AI 研究一大弱点便是视觉知识研究的不足. 视觉知识的研究与运用是 AI 2.0（Pan, 2016）的一个重要发展方向.

图 1 三四岁儿童已能识别并使用不同杯子，但未能以语言确述之

2 基本问题 2：视觉识别

从 AI 早期开始，模式识别便是其中一个最重要的研究领域，其中图象和视频识别是发展最快的方向.

曾使用基于数字图象处理技术的图象识别技术，是一种将局部特征综合为整体对象的方法. 近来，深度学习以端到端方式提供了另一方法：用大量标识的图像训练出深度神经网络模型（DNN）用于图象识别，显著提高正确率，已获广泛应用.

DNN 具有的优点是：（1）可从被标识的样本数据中通过学习自动获得模型知识；（2）可用于非符号数据的识别，如图像与语音识别.

但 DNN 也有缺点：（1）难以解释；（2）不可推理；（3）需要大量被标识的数据来训练网络参数，从而获得知识.

值得注意的是，与 DNN 方法不尽相同，人类在工作记忆中进行视觉识别时，不仅分析视网膜感知后传入短期记忆中的数据，而且激活了工作记忆处理过程中所需长期记忆中的相关心象，即视觉知识. 正因为如此，人类在完成视觉识别任务时往往只需少量数据，而且可解释，也可推理.

因此在视觉识别中，不但使用数据，而且协同使用视觉知识，形成数据驱动和视觉知识指导的计算范式是重要的研究方向.

3 基本问题 3：视觉形象思维模拟

形象思维是人类在设计、创意和问题求解时重要的智能行为. 模拟形象思维，需要如下操作：（1）视觉形象的物理变化，如几何变换、时空变换、场景变换，比较、预测、分解与装配等；（2）视觉形象的生物变化，如运动、生成、互动等；（3）视觉形象的想象变化，如创意与设计新产品（《西游记》《阿凡达》《狮子王》《小飞象》等）中的各种想象性操作.

视觉形象思维模拟在计算机辅助设计（CAD）、计算机动画、游戏、儿童教育和数字媒体创意等领域应用十分广泛. 按数字媒体的不同，可分为 3 类：（1）从文本生成视觉形象. 如，给出一段文字描写，自动生成一个图形图象的背景. 又如，给出一段评价，自动修改产品的设计. （2）从一种视觉形象变换为另一种形象. 如，大闹天宫中将孙悟空从猴子自动变为一座庙. 又如，浙江大学现代工业设计研究所的研究人员，将正方形中的标准篆字变为印章设计中汉印风格的篆字（图 2）. （3）从视觉形象生成文本. 如，给一张图片或一段视频赋予一个标题或生成一段语言描述，并且分类. 短视频内容的文本描述自动生成，现已用于网上销售的商品精准推荐（Zhang SY et al., 2020）.

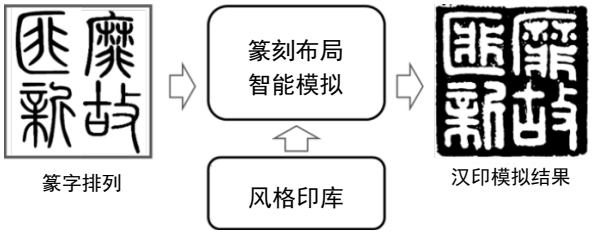

图 2 形象模拟 AI 技术用于篆刻布局

计算机图形学已储备很多基础技术，但有待与 AI 打通. 一旦实现，有望形成新一代设计软件的基础.

4 基本问题 4：视觉知识的学习

计算机视觉（CV）已经体现视觉对象形体重构的重要性，并积累了很多成果，如 3D 扫描重构形体、多相机重构形体、基于视频重构形体等等. 形体重构是 CV 和 CG 的桥梁.

然而，视觉知识学习则要将目标从视觉形状的重建任务提升到视觉知识概念和命题的重建，这就需要对现有计算机视觉技术做进一步研究：不仅要重建 3D 形状，而且要重构 3D 形状的概念结构与层次结构.

在此基础上，有望发展出视觉知识的自动学习手段. 当前的场景图（Xu et al., 2017; Zellers et al., 2018）研究是向视觉知识自动学习前进的一个合适的中间方法.

为此，特别需要当今人工智能、计算机图形学和计算机视觉 3 个领域的研究者们联手研究视觉知识的自动学习.

5 基本问题 5：多重知识表达（Pan, 2020）

人脑中的知识往往是通过多重表达来描述. 所以，在 AI 2.0 中的知识应有多种表达方式. 列出 3 种知识的表达与处理方法如下：（1）知识的言语表达. 其特点是使用符号数据，因此结构清晰，语义可理解，知识可推理. 其典型例子如语义网络、知识图谱（Zhang NY et al., 2020）. 目前，此类知识的获取正在从人工构造向自动抽取过渡（Tang et al., 2018）. （2）知识的深度神经网络表达. 其特点是适用于图像、音频等非结构化数据的分类与识别，缺点是语义解释困难. 其典型例子如 DNN. 目前，此类知识的获取正在从人工标注的监督学习向无监督学习发展（Brown et al., 2020）.（3）知识的形象表达. 其特点是适用于图形、动画等描述形状、空间、运动的数据. 这一类知识结构清晰、语义可解释、知识可推演. 其典型例子如视觉知识. 目前，此类知识的获取与利用仍是一个亟需研究和发展的方向.

AI 的这 3 种知识表达符合人类记忆中的 3 种不同但相通的内容，现说明如下：（1）知识图谱——语义的记忆内容，宜用于字符类信息的检索与推理；（2）视觉知识——视觉情景的记忆内容，宜用于视觉形象类信息的时空推演与可视化；（3）深度神经网络——感知的记忆内容，宜用于对原始数据中的模式通过逐层抽象进行学习，进而分类.

其中（1）和（2）与人类长期记忆中两大内容——言语和心象的编码方式——相对应. 其中（3）与人类短期记忆中的感知内容相对应.

AI 2.0 要令多种知识表达相通使用，这就是多重知识表达. 它将形成跨媒体智能（Zhuang et al., 2017）和大数据智能的技术基础.

从视觉智能的 5 大问题分析可知，问题 1、2、4 的解决有较好基础，问题 3、5 尚需多领域学者协力攻关. 由此可见，视觉知识及跨媒体知识表达是关键所在.

6 小结

由上述分析可知，视觉知识的独特优点是具有形象的综合生成能力、时空演化能力和形象显示能力. 这些正是字符知识和 DNN 所缺乏的. AI 与 CAD/CG/VC 技术联合将为 AI 在创造、预测和人机融合等方面的新发展提供重要的基础动力.

视觉知识和多重知识表达的研究是发展新的视觉智能的关键，也是促进 AI 2.0 取得重要突破的关键理论与技术.

这是一块荒芜、寒湿而肥沃的“北大荒”，也是一块充满希望、值得多学科合作勇探的“无人区”.

致谢

感谢孙凌云、肖俊、张克俊教授以及邓晃煌先生为本文提供富有价值的资料.

References

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. https://arxiv.org/abs/2005.14165

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

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

Pan YH, 2020. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216-217. https://doi.org/10.1016/j.eng.2019.12.011

Tang SL, Zhang Q, Zheng TP, et al., 2018. Two step joint model for drug drug interaction extraction. https://arxiv.org/abs/2008.12704

Xu DF, Zhu YK, Choy CB, et al., 2017. Scene graph generation by iterative message passing. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5410-5419.

Zellers R, Yatskar M, Thomson S, et al., 2018. Neural motifs: scene graph parsing with global context. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5831-5840.

Zhang NY, Deng SM, Zhang W, et al., 2020. Relation adversarial network for low resource knowledge graph completion. Proc Web Conf, p.1-12. https://doi.org/10.1145/3366423.3380089

Zhang SY, Tan ZQ, Zhou Z, et al., 2020. Comprehensive information integration modeling framework for video titling. Proc SIGKDD Int Conf on Knowledge Discovery & Data Mining, p.2744-2754. https://doi.org/10.1145/3394486.3403325

Zhuang YT, Jain R, Gao W, et al., 2017. Panel: cross-media intelligence. Proc 25th ACM Int Conf on Multimedia, p.1173. https://doi.org/10.1145/3123266.3133336