机器与人类视觉能力的差距（1）

> 本文属于个人观点，跟本人在职公司的立场无关。由于最近 GitHub 服务器在国内访问速度严重变慢，虽然经过大幅度压缩尺寸，文中的图片仍然可能需要比较长时间才能加载。这篇文章揭示了 AI 领域重要的谬误和不实宣传，为了阻止愚昧的蔓延，我鼓励大家转发这篇文章和它的后续，转发时只需要注明作者和出处就行。

很多人以为人工智能就快实现了，往往是因为他们混淆了“识别”和“理解”。现在所谓的“人工智能”都是在做识别：语音识别，图像识别，而真正的智能是需要理解能力的。我们离理解有多远呢？恐怕真正的工作根本就没开始。

很长时间以来，我都在思索理解与识别的差别。理解与识别是很不一样的，却总是被人混为一谈。我深刻的明白理解的重要性，可是我发现很少有其他人知道“理解”是什么。AI 领域因为混淆了识别和理解，一直以来处于混沌之中。

最近因为图像识别等领域有了比较大的进展，人们对 AI 产生了很多科幻似的，盲目的信心，出现了自 1980 年代以来最大的一次“AI 热”。很多人以为 AI 真的要实现了，被各大公司鼓吹的“黑科技”冲昏了头脑，却看不到现有的 AI 方法与人类智能之间的巨大鸿沟。所以下面我想介绍一下我所领悟到的机器和人类在视觉能力方面的差距，希望一些人看到之后，能够再次拥有冷静的头脑。

在之前一篇文章《人工智能的局限性》中，我已经阐述了对自然语言处理领域误区的看法。当时因为对计算机视觉方面了解不多，所以没有包含视觉方面的内容。熟悉了机器视觉的各种做法之后，我想在这篇文章里详述一下视觉方面的内容。这两篇文章加在一起，可以说概括了我对 AI 语言和视觉两个方面的领悟。

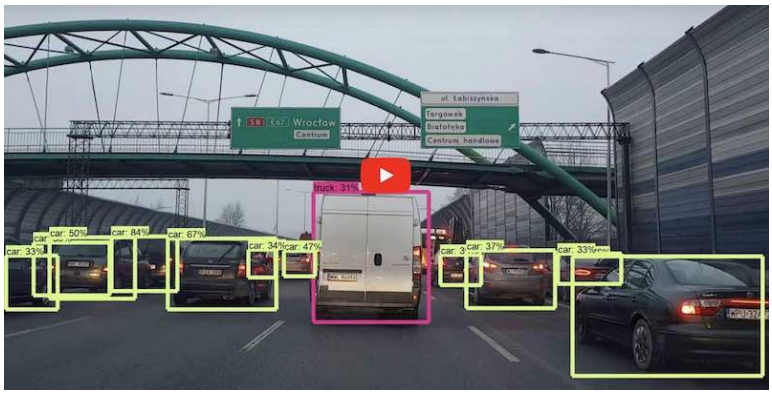
“图像识别”和“视觉理解”的差别

对于视觉，AI 领域混淆了“图像识别”和“视觉理解”。现在热门的所谓 “AI” 都是“图像识别”，而动物的视觉系统具有强大的“视觉理解”。视觉理解和图像识别有着本质的不同。

深度学习视觉模型（CNN一类的）只是从大量数据拟合出从“像素=>名字”的函数。它也许能从一堆像素猜出图中物体的“名字”，但它却不知道那个物体“是什么”，无法对物体进行操作。注意我是特意使用了“猜”这个字，因为它真的是在猜，而不像人一样准确的知道。

“图像识别”跟“语音识别”处于同样的级别，停留在语法（字面）层面，而没有接触到“语义”。语音识别是“语音=>文字”的转换，而图像识别则是“图像=>文字”的转换。两者都输出文字，而“文字”跟“理解”处于两个不同的层面。文字是表面的符号，你得理解了它才会有意义。

怎样才算是“理解了物体”呢？至少，你得知道它是什么形状的，有哪些组成部分，各部分的位置和边界在哪里，大概是什么材料做成的，有什么性质。这样你才能有效的对它采取行动，达到需要的效果。否则这个物体只是一个方框上面加个标签，不能精确地进行判断和操作。



想想面对各种日常事物的时候，你的脑子里出现的是它们的名字吗？比如你拿起刀准备切水果，旁边没有人跟你说话，你的脑子里出现了“刀”这个字吗？一般是没有的。你的脑子里出现的不是名字，而是“常识”。常识不是文字，而是一种抽象而具体的数据。

你知道这是一把刀，可是你的头脑提取的不是“刀”这个字，而是刀“是什么”。你的视觉系统告诉你它的结构是什么样的。你知道它是金属做的，你看到刀尖，刀刃，刀把，它也许是折叠的。经验告诉你，刀刃是锋利的可以切东西的部分，碰到可能会受伤，刀把是可以拿的地方。如果刀是折起来的，你得先把它翻开，那么你从哪一头动手才能把它翻开，它的轴在哪里？

你顺利拿起刀，开始切水果。可是你的头脑里仍然没有出现“刀”这个字，也没有“刀刃”，“刀把”之类的词。在切水果的同时，你大脑的“语言中心”可能在哼一首最近喜欢的歌词，它跟刀没有任何关系。语言只是与其他人沟通的时候需要的工具，自己做事的时候我们并不需要语言。完成切水果的动作，你需要的是由视觉产生的对物体结构的理解，而不是语言。

你不需要知道一个物品叫什么名字就能正确使用它。同样的，光是知道一个物品的名字，并不能帮助你使用它。看到一个物体，如果脑子里首先出现的是它的名字，那么你肯定是很愚钝的人，无法料理自己的生活。现在的“机器视觉”基本就是那样的。机器也许能得出图片上物体的名字，却不知道它是什么，无法操作它。

试想一下，一个不能理解物体结构的机器人，它只会使用图像识别技术，在你的头上识别出一个个的区域，标注为“额头”，“头发”，“耳朵”…… 你敢让它给你理发吗?

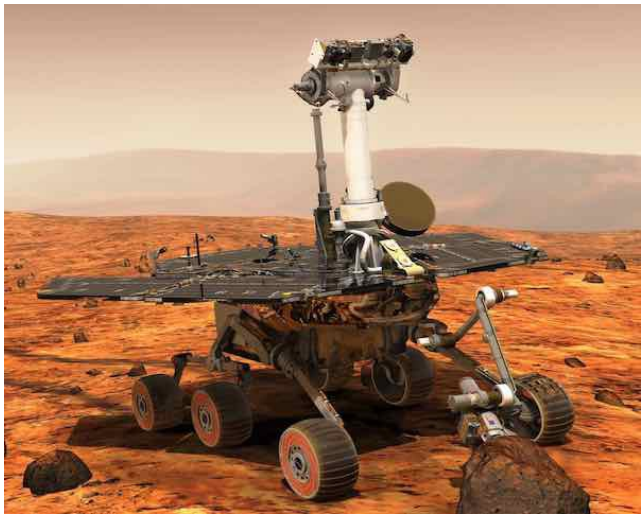
这就是我所谓的“视觉理解”与“图像识别”的差别。你会意识到，这种差别是巨大的。

视觉识别不能缺少理解

如果我们降低标准，只要求识别出物体的名字，那么以像素为基础的图像识别，比如卷积神经网络（CNN），也是没法像人一样准确识别物体的。人识别物体并不是像神经网络那样的“拍照，识别”两节拍动作，而是一个动态的，连续的过程：观察，理解，观察，理解，观察，理解……

感官接受信息，中间穿插着理解，理解反过来又控制着观察的方向和顺序。理解穿插在了识别物体的过程中，“观察/理解”成为不可分割的整体。人看到物体的一部分，理解了那是什么，然后继续观察它周围是什么，反复这个过程，最后才判断出物体是什么。机器在识别的过程中没有理解的成分存在，这就是为什么机器在图像识别能力上无法与人类匹敌。

这个“观察/理解”的过程发生的如此之快，眨眼间就完成了，以至于很多人都没察觉到其中“理解成分”的存在。所以我们现在放慢这个过程，来一个慢镜头特写，看看到底发生了什么。假设你从来没见过下面这个东西，你知道它是什么吗?



一个从没见过这东西的人，也会知道这是个“车”。为什么呢？因为它有轮子。为什么你知道那是轮子呢？仔细一想，因为它是圆的，中间有轴，所以好像能在地面上滚动。为什么你知道那是“轴”呢？我就不继续折腾你了，自己想一下吧。所有这些分析都是“视觉理解”所产生的，而这些理解依赖于你一生积累的经验，也就是我所谓的“常识”。

其实为了识别这个东西，你并不需要分析这么多。你之所以做这些分析，是因为另一个人问你“你怎么知道的？”人识别物体靠的是所谓“直觉”。一看到这个图片，你的脑子里自然产生了一个 3D 模型。一瞬间之后，你意识到这个模型符合“车”的机械运动原理，因为你以前看见过汽车，火车，拖拉机…… 你的脑子里浮现出这东西可能的运动镜头，你仿佛看到它随着轮子在动。你甚至看到其中一个轮子压到岩石，随着连杆抬了起来，而整个车仍然保持平衡，没有反倒，所以这车也许能对付崎岖的野外环境。

这里有一个容易忽视的要点，那就是轮子的轴必须和车体连在一起。如果轮子跟车体没有连接，或者位置不对，看起来无法带着车体一起运动，人都是知道的。这种轮轴与车身的连接关系，属于一种叫“拓扑”（topology）的概念。

拓扑学是一门难度挺高的数学分支，但人似乎天生就理解某些浅显的拓扑概念。实际上似乎高等动物都或多或少理解一些拓扑概念，它们一看就知道哪些东西是连在一起的，哪些是分开的。捕猎的动物都知道，猎物的尾巴是跟它们身体连在一起的，所以咬住它们的尾巴就能抓住它们。

拓扑学还有一个重要的概念，那就是“洞”。聪明一点的动物基本上都理解“洞”的概念。很显然老鼠，兔子等穴居动物必须理解洞是什么。它们的天敌，猫科动物等，也理解洞是什么。如果我拿一个纸箱给我的猫玩，我在上面挖一个洞，等他钻进去，他是不会进去的。我必须在上面挖两个洞，他才会进去。为什么呢？因为他知道，要是箱子上面只有一个洞，要是他进去之后洞被堵上，他就出不来了！

机器如何才能理解洞这个概念呢？它如何理解“连续”?

总之，人看到物体，他看到的是一个 3D 模型，他理解其中的拓扑关系和几何性质，所以一个人遇到前所未见的物体，他也能知道它大概是什么，推断出如何使用它。理解使得人可以非常准确地识别物体。没有理解能力的机器是做不到这一点的。

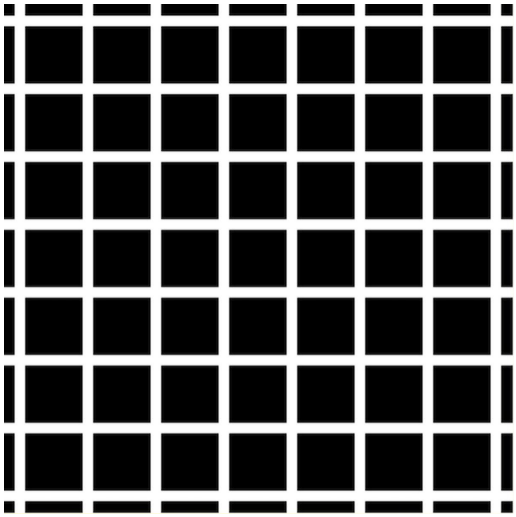
人的视觉系统与机器的差别

人的眼睛与摄像头有着本质的差异。眼睛的视网膜中央非常小的一块区域叫做“fovea”，里面有密度非常高的感光细胞，而其它部分感光细胞少很多，是模糊的。可是眼睛是会转动的，它被脑神经控制，敏捷地跟踪着感兴趣的部分：线条，平面，立体结构…… 人的视觉系统能够精确地理解物体的形状，理解拓扑，而且这些都是 3D 的。人脑看到的不是像素，而是一个 3D 拓扑模型。

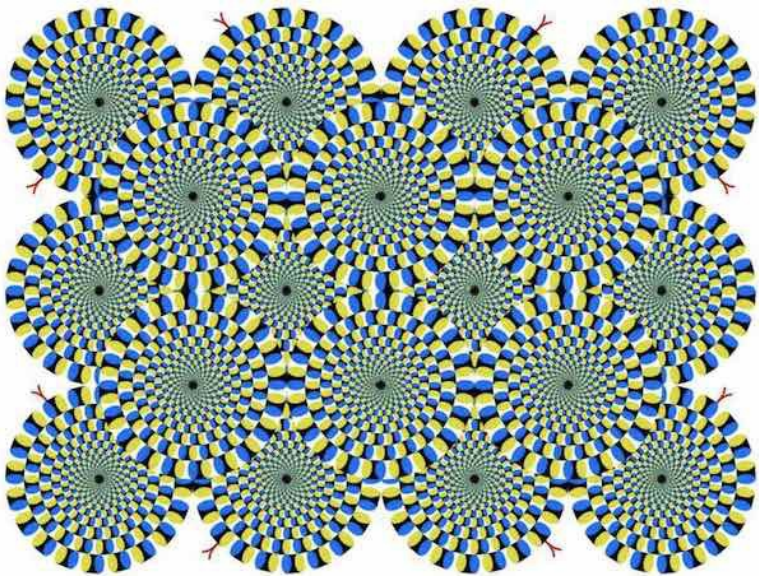
眼睛观察的顺序，不是一行一行从上往下把每个“像素”都记下来，做成 6000x4000 像素的图片，而是聚焦在重点上。它可以沿着直线，也可以沿着弧线观察，可以转着圈，也可以跳来跳去的。人脑通过自己的理解能力，控制着眼睛的运动，让它去观察所需要的重点。由于视网膜中央分辨率极高，所以人脑可以得到精度非常高的信息。然而由于不是每个地方都看的那么仔细，所以眼睛采集的信息量可能不大，人脑需要处理的信息也不会很多。

人的视觉系统能理解点，线，面的概念，理解物体的表面是连续的还是有洞，是凹陷的还是凸起的，分得清里和外，远和近，上下左右…… 他能理解物体的表面是什么质地，如果用手去拿会有什么样的反应。他能想象出物体的背面大概是什么样子，他能在头脑中旋转或者扭曲物体的模型。如果物体中间有缺损，他甚至能猜出那位置之前什么样子。

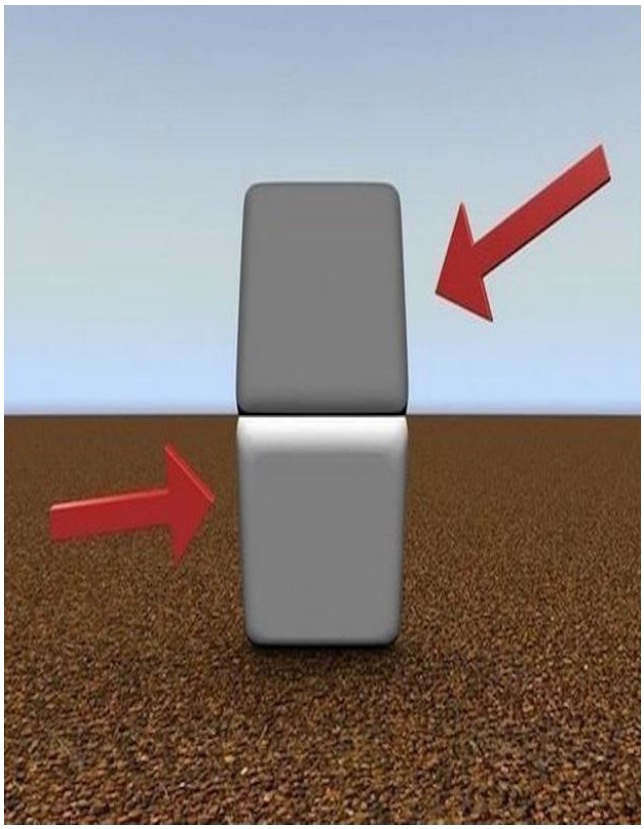
人的视觉系统比摄像头有趣的多。很多人都看过“光学幻觉”（optical illusion）的图片，它们从一个角度揭示了人的视觉系统背后在做什么。比如下图本来是一个静态的图片，可是你会感觉有很多暗点在白线的交叉处，但如果你仔细看某一个交叉处，暗点却又不见了。这个幻觉很经典，被叫做 Herman grid，在神经科学界被广泛研究。稍后我还会提到这个东西。



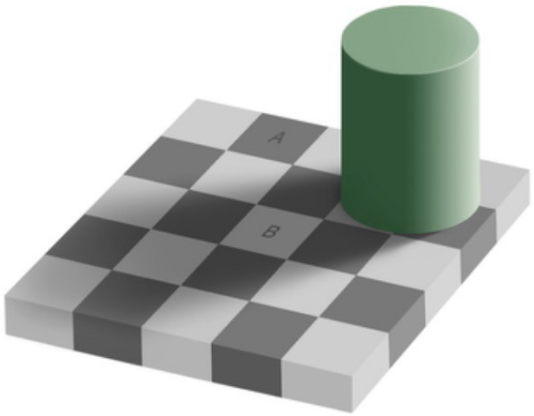
本来是静态图片，你却感觉它在转。



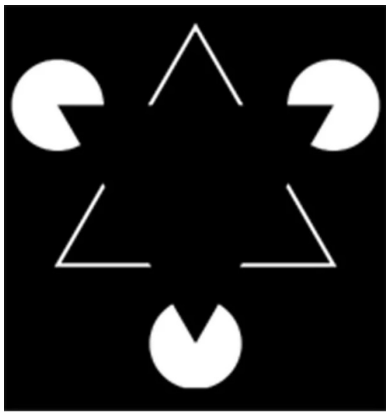
本来上下两块东西是一样的颜色，可是看起来下面的颜色却要浅一些。如果你用手指挡住中间的高亮部分，就会发现上下两块的颜色其实是一样的。



另一个类似的幻觉，是著名的“Abelson 棋盘幻觉”。图中 A 和 B 两个棋盘格子的颜色是一样的，你却觉得 A 是黑色，而 B 是白色。不信的话你可以用软件把这两块格子从图片上切下来，挨在一起对比一下。如果你好奇这是为什么，可以参考这篇文章。



在下图里，你会觉得看见了一个黑色的倒三角形，可是其实它并不存在。



很多的光学幻觉都说明人的视觉系统不是简单的摄像头一样的东西，它具有某些特殊功能。这些特殊功能和机制导致了这些幻觉。这使得人类视觉不同于机器，使得人能够提取出物体的结构信息，而不是只看到像素。

提取物体的拓扑结构特征，这就是为什么人可以理解抽象画，漫画，玩具。虽然世界上没有猫和老鼠长那个样子，一个从来没看过《猫和老鼠》动画片的小孩，却知道这是一只猫和一只老鼠，后面有个房子。你试试让一个没有拿《猫和老鼠》剧照训练过的深度学习模型来识别这幅图？

更加抽象的玩具，人也能识别出它们是哪些人物。头和四肢都变成了方的，居然还是觉得很“像”。你不觉得这很神奇吗？

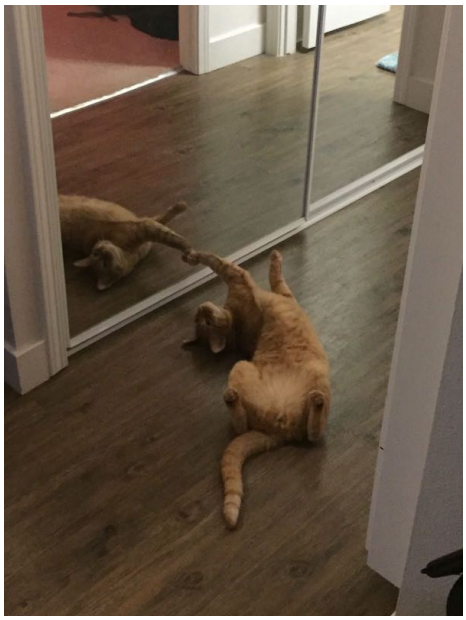
人脑理解“拓扑”的概念，这使得人能够不受具体像素干扰而正确处理各种物体。对拓扑结构的理解使得人对物体的识别非常准确，甚至可以在信息不完整，模糊，扭曲的情况下工作，在恶劣的天气环境下，有反光，有影子的情况下也能识别物体。

说到反光，你有想过机器要如何才能识别出场景里有一面镜子或者玻璃吗？如果场景中有反光的物体，比如镜子，平静的水面，镀铬的物品，神经网络（CNN）那种依靠像素滤镜训练出来的函数还会有用吗？要知道它们看到的像素，可能有一大片是通过镜面反射形成的，所以无法通过局部的纹理识别出这种情况来。

这是个现实的问题。自动车或者机器人要如何知道前面的路面上有积水或者结冰了？它们要如何知道从水面反射过来的镜像不是真实的物体？比如，它们如何知道下图里路面上的倒影不是真正的树呢？要知道，倒影的像素纹理，跟真实的场景可能是非常相似的。

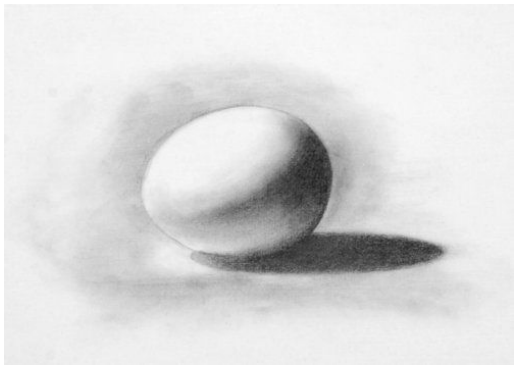
人是通过对光的理解，各种常识来识别镜子，玻璃，地上的水和冰的存在。一个不理解光和水的性质的机器，它能察觉这些东西的存在吗？靠像素分析能知道这些？要知道，这些东西在某些地方出现，可以是致命的危险。

很有趣的事情，理解光线的反射和折射，似乎已经固化到了每个动物的视觉系统里面。我观察到这一点，是因为我的卧室和客厅之间的橱柜门上有两面大镜子。我的猫在卧室里，能够从镜子里看见我在客厅拿着逗猫绳。他冲过来的时候却不会撞到镜子上面，而是出了卧室门立马转一个角度，冲向我的方向。我每次看到他敏捷的动作都会思考，他是如何知道镜子的存在呢？他是如何知道镜子里的猫就是他自己，而不是另一只猫？



人脑会构造事物的 3D 模型

说了光，再来说影吧。画过素描的人都知道，开头勾勒出的轮廓是没有立体感的，然后你往恰当的位置加一些阴影，就有了立体感。所以动物的视觉系统里存在对影子的分析处理，而且这种功能我们似乎从来没需要学习，生下来就有。“立体视觉”是如此强烈的固化到了我们的头脑里，一旦产生了立体感，你就很难再看见平面的像素。

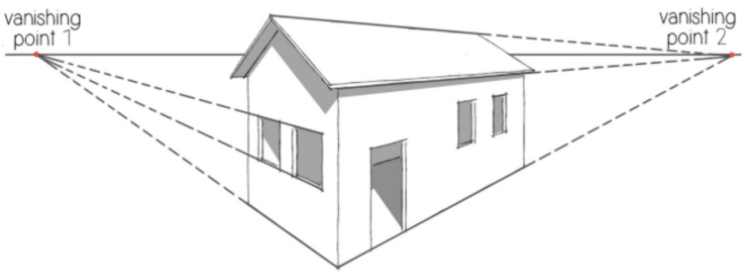


靠着光和影的组合，人和动物能得到很多信息。比如上图，我们不但看得出这是一个立体的鸡蛋，而且能推断出鸡蛋下面是一个平面，可能是一张桌子，因为有阴影投在了上面。

神经网络知道什么是影子吗？它如何知道影子不是实际存在的物体呢？它能从影子得到有用的信息吗？

神经网络根本不知道影子是什么。早就有人发现，Tesla 基于图像识别的 Autopilot 系统会被阴影所迷惑，以为路面上的树影是一个障碍物，试图避开它，却差点撞上迎面来的车。我在很早的一篇文章已经谈过这个问题。

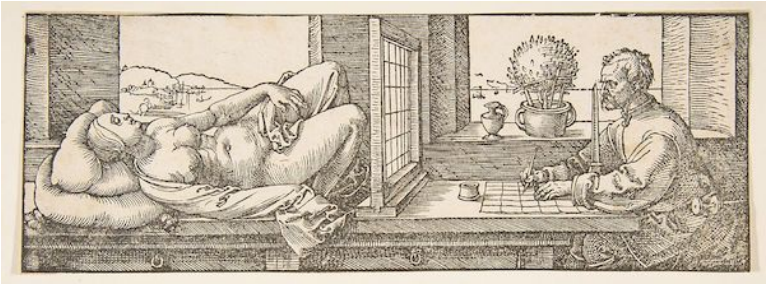
再来一个关于绘画的话题。学画的初期，很多人都发现画“透视”特别困难。所谓透视就是“近大远小”。本来房子的几堵墙都是长方形，是一样高的，可是你得把远的那一边画短一些，而且相关部分的比例都要画对，就像照片上那样，所以墙就成了梯形的。房顶，窗户等，也全都得做相应的调整。你得这样画，看画的人才会感觉是对的，不然就会感觉哪里不对劲，不真实。



这件事真的很难，大部分人（包括我）一辈子都没学会画透视。虽然拿起笔来量一下，我确实看到远的那一边要

短一些，可是我的脑子似乎会“自动纠错”，让我认为它们都是一样长的。所以要是光靠眼睛徒手作画，我会把那些边都画成一样长。我似乎永远学不会画画！

画透视是如此困难的事情，以至于 16 世纪的德国画家丢勒为此设计了一种专门的设备。



你可能没有想到，这个使得我们学画困难的罪魁祸首，其实是人类视觉系统的一项重要功能，它帮助我们理解身边的环境。虽然眼睛看到的物体是近大远小，可是人脑会自动调整它们在你“头脑里的长度”，所以你知道它们是一样长的。

这也许就是为什么人能从近大远小的光学成像还原出正确的 3D 模型。在你头脑中的模型里面，房子的几堵墙是一样高的，就像它们在现实中的情况一样。有了准确的 3D 模型，人才能正确地控制自己在房子周围的运动。

这种导致我们学画困难的“3D 自动纠错”功能，似乎固化到了每个人，每个高等动物的视觉系统里。我们并不需要学习就有这种能力，它一直都在起作用。反倒是我们要想“关掉”这个功能的时候，需要付出非常多的努力！

为什么人想要画出透视效果那么困难呢？因为一般人画画，都不是在画他们头上那两只眼睛看到的东西，而是在画他们的“心之眼”（mind’s eye）看到的东西——他们头脑中的那个 3D 模型。这个 3D 模型是跟现实“同构”的，模型里房子的墙壁都是一样高的，他们画出来也是一样高的，所以就画错了。只有经过专业训练的画家，才有能力关闭“心之眼”，直接画出眼睛看到的东西。

我猜想，每一种高等动物的视觉系统都有类似的机制，使得它们从光学成像“重构”出与现实同构的 3D 模型。缺乏 3D 建模能力的机器，是无法准确理解看到的物体的。

现在很多自动驾驶车用激光雷达构造 3D 模型，可是相对于人类视觉形成的模型，真是太粗糙了。激光雷达靠主动发射激光，产生一个扫描后的“点云”，分辨率很低，只能形成一个粗糙的 3D 轮廓，无法识别物体，也无法理解它的结构。我们应该好好思考一下，为什么人仅靠被动接收光线就能构造出如此精密的 3D 模型，理解物体的结构，而且能精确地控制自己的动作来操作这些物体。

现在的深度学习模型都是基于像素的，没有抽象能力，不能构造 3D 拓扑模型，甚至连位置关系都分不清楚。缺乏人类视觉系统的这种“结构理解”能力，可能就是为什么深度学习模型需要那么多的数据，那么多的计算，才勉强能得出物体的名字。而小孩子识别物体根本不需要那么多数据和计算，看一两次就知道这东西是什么了。

人脑提取了物体的要素，所以很多信息都可以忽略了，所以人需要处理的数据量，可能比深度学习模型小很多。深度学习领域盲目地强调提高算力，制造出越来越大规模的计算芯片，GPU，TPU…… 可是大家想过人脑到底有多大计算能力吗？它可能并不需要很多计算。

从上面的各种现象，我们也许已经看明白了，人类视觉系统是很神奇的。现有的机器视觉研究并没有理解人类视觉的这些能力是怎么实现的。在接下来的续集中我们会详细的看清楚，AI 领域到底理解多少人类神经系统的构造。

请看下一篇：机器与人类视觉能力的差距（2）

（这个系列的文章包含了很多独到的见解。如果你觉得有帮助可以付费支持。）