格灵深瞳邓亚峰：如何让机器看得更真更远？

马春杰

视觉在人类获取信息的方式中所占比例大约是百分之八十几，处于绝对领先地位。

### 1 计算机视觉的研究目标和应用趋势

计算机视觉系统的输入是一系列的传感器，如visual sensor、RGB sensor等。计算机视觉的研究目标是：**输出对这个世界的理解**。

应用趋势有很多，包括安防、工业检测、智慧医疗、智能驾驶等。

### 2 面临的挑战与机遇

#### 2.1 挑战

ILSVRC2016上的Object Detection的正确率只有66.3%，这只能应用于一些准确率要求不高的场景。

#### 2.2 机遇

* 整个计算机领域迎来了前所未有的关注和投资热潮；
* 计算机视觉领域的应用呈现了爆发式的增长态势；
* 从世界范围来讲，华人在计算机领域处于一个领先地位，这使得我们可以和美国去争夺人工智能的制高点。

### 3 如何让机器看的更真更远

过去的特征和学习是分开的，在深度学习时代，特征和学习被融合起来，这样做的好处就是，特征的表示不必完全依赖专家的选择，此时的特征完全是基于数据驱动的，并且是一层一层的非线性表示，所以能够很方便的得到对一个数据非常强的表示。

深度学习成功的三个因素：**深度学习算法**、**大规模数据**以及**硬件平台**。

如何使深度学习与数据形成一种良性循环呢，首先需要一些初始的数据，数据量大约在几十万到几千万级别，然后经过深度学习之后，会得到一个模型，将这个模型应用到系统中去，系统再**反馈**那些最应该标注的数据，然后再使用这些应该标注的数据当做训练数据**反馈**到系统中，这样就形成了一种良性循环。

对于图像识别而言，最常用的是RGB传感器，对于只利用RGB信息的方法，我们称作2D方法，还有一种传感器增加了一个Depth的信号，称为RGBD传感器，基于这种传感器的方法称为3D方法。

格灵深瞳的3D方法：

* 首先基于depth输入做前景提取，得到潜在的目标区域
* 然后做投影变换，把潜在的目标投影到水平面
* 结合RGB的方法产生一些proposal
* 使用深度学习的方法做proposal的后处理过滤

这样就得到了一个快速且效果好的多目标检测器。

#### 感知问题

“感”是获取图像，“知”是处理图像。

对于获取图像，作者用单反相机做实验，人在距离单反十米左右，人脸分辨率大概是36x36，二十米左右，分辨率是18x18，三十米左右，分辨率为12x12。对于普通的1080P的监控摄像头，基本上三五米之内可以识别，再远就不行了。

使用一个比较经典的VGG模型对人脸进行表示，研究发现，当人脸的分辨率达到50的时候，准确率大概只有80%，如果是25像素，准确率为70%，当分辨率达到100像素以上，准确率可以达到95%左右。这表明，对人脸识别来说，图像分辨率是非常重要的。

所以提高精度的一个方法就是**提高目标的分辨率**。但实际上**宽和远是不可兼得的**。另一种方法是提高摄像机的分辨率，但是提高摄像机的分辨率会造成数据量的增加，使得成本增加，并且对网络传输和存储也会造成很大的压力。

格灵深瞳开发了一种人眼相机，模拟人眼去采集和理解图像，最终达到了有效100倍的提升。

### 计算机视觉大规模应用的必经之路

决定能否大规模应用的两个因素，第一是准确率，第二是成本。

从技术层面讲，必经的三个阶段是：

1. 云的方式。类似于服务器，将视频传回服务器端，在服务器端进行处理运算，这样就能获得大量的数据，有利于算法的迭代；
2. 云+端的方式。如果全部用云的方式，带宽消耗比较大，所以可以结合端，通过端来做一些运算量比较小的事。
3. 芯片的方式。不过这种方式需要等到算法成熟之后才可以，不然的话，算法更新之后，芯片再更新的话消耗比较大。

**仅仅有算法不行，一定要跟产业结合起来。**