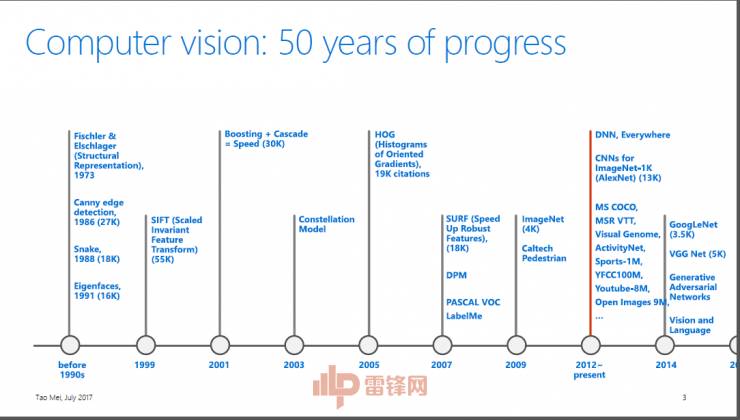
深度学习近年来取得了历史性的进步，可以在静止的图像中识别物体，表现不输给人类，但是计算机理解视频的空间和时间仍然是一个未解的问题。因为照片只是一张静态图片，但一段视频则是“动作的叙述”，透露的信息包括了三维几何、材料性质、物体持久性、重力等。

**视频自动标签：**

微软亚洲研究院目前把这个技术用在了聊天机器人的自动评价功能里，例如微软小冰，当用户上传视频给小冰，它会夸赞对方。在这个技术上线一个月后，小冰在某视频网站上的粉丝数涨了60%。当然，小冰现在还可以根据图片内容写现代诗，将来我们希望小冰能够根据视频来写诗。

**视频生成：**

“我们也可以将Video进行编辑，加上滤镜，或是做风格的转换，把自然的Video变得非常卡通。视频中的人物分割出来可以放到另外一个虚拟的场景里面去。你可以想象，当两个人在异地谈恋爱的时候，我们能够给他一个房间，让他们在同一个房间里、在星空下、在安静湖面上的一艘小船上进行聊天。另外，我们也可以提供storytelling的服务，让原始的、没有经过任何编辑和处理的image、video集合变成一段非常吸引人的、有一定设计感和视觉感的故事，这段视频demo就是机器自动产生的效果。加上人工的处理，视频就可以变得更加时尚。”



从那开始，所有视觉的东西都在用CNN，代表性的有GoogLeNet，AlexNet等等，我们的任务也会越来越多，越来越有挑战，比如现在正在做的从图片中生成语言，不仅要在图片或视频中打上一些标签，还要把这些标签变成能用自然语言描述的一句话。

我们今天讲的是视频理解，如果从理解一个像素或理解一个图片或视频来说，可以把这个理解问题分成几个层次。最难的是需要理解图片或视频里面每个像素代表什么标签。再往上是我们关心每一个物体在什么位置、属于什么类别。第三部分是不关心这个物体在什么地方，你给我一图片或者视频，我就知道这个图片或者视频里面有什么标签。再往前走一步，比如说我给你一个图片，要求不仅要生成单独的标签，还要看你能不能生成一个非常自然的语言来描述这个图片。再往上，我给你一个图片，能不能给我一个故事，比如说现在机器能不能产生这样一个故事。

今天从另外一个角度来看视频内容的产生、编辑、管理会经历哪些过程，有哪些技术来支撑，我们从Creation（创作）到Curation（处理）、到Consumption（消费）的顺序来讲。

比如说来了一个运动视频，通过智能分析知道这个运动视频里面哪个部分最应该看，这是它的重点。

*另一个话题是视频生成，今天我给你一段文字，你给我生成一个新的视频，这个事情听起来天方夜谭，但是值得挑战。我告诉你生成一个视频，也就是一个数字8在上面不停的游动。另外给你数字6和0，能不能让数字6和0在这里面游动，这个事情是非常难的。最近我们做了研究，发现可以做一些简单的事情，比如说一个人在烤牛肉。实际上这还是很难的，因为我们生成的视频准确性非常低，所以这是非常难的事情。*

当有了视频之后要做的事情是**给视频打标签**，至今为止可以打上1000个的静态标签，你有了这些静态的标签就可以设置到内容里面去。比如说视频里面出现一个桥，这个桥的位置在哪里。比如说一些运动，这是我们能够识别出来的运动，左边都是运动的视频，右边是我们日常生活中的一些行为。有两个动作最难识别，一个是跳跃，一个是三级跳，但是我们现在已经可以区分出这些非常细微的差别。

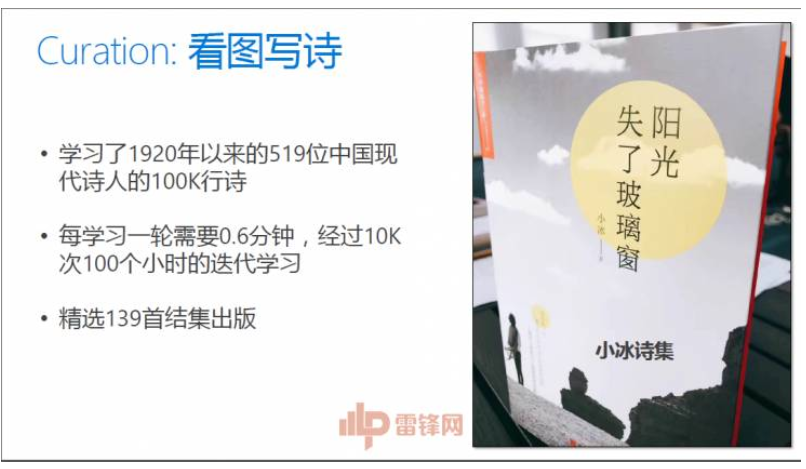
这是我们今天讲的唯一的一个技术性的部分。我们最近做的一个非常好的工作，就是可以做深层次的网络，我们可以通过一些方式使得深层次的网络是可行的。比如说这个到现在可以做到152层，也可以做到1001层，性能超过了任何网络。我们能不能从这个网站的图片扩展到视频？我把二维的卷积盒变成三维的，当它卷积的时候是沿着X、Y和T这个方向卷积的。C3D模型是可以做到13层，它非常复杂。我们有一个想法，把它进行分解，一个是可以找出这个物体的数量，使的这个过程可行，另外还可以在图像上进行处理。我们做了很多工作，比如说这个视频是一个太极的动作，我们通过P3D可以找出来4个点，这个已经非常了不起了。

**我们可以很精确的告诉你这个视频中每一个关节是怎么运动的（见PPT），比如说我今天做一个智能的健身教练，可以把你的动作进行分解，告诉你哪个动作是不准确的。**

***还有一个是Video captioning（视频说明），给你一个视频，能不能生成一句话来描述这个视频。以前我们说这个视频是一个舞蹈，现在可以告诉你说这是一个什么舞蹈。***



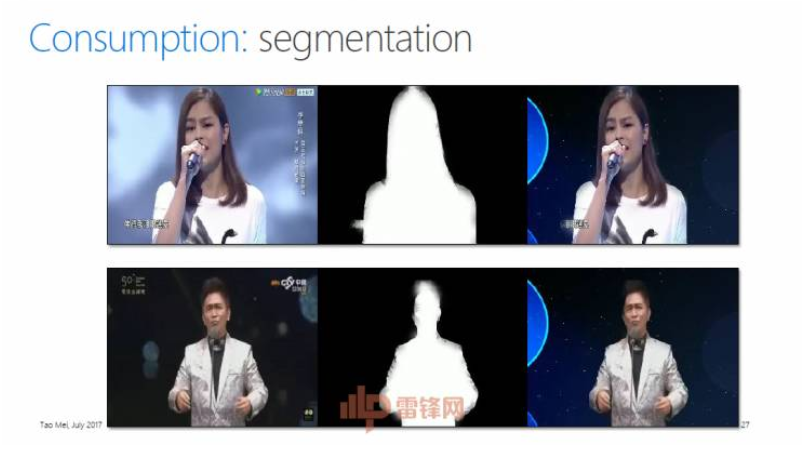
**小冰还可以写诗，最近我们发表了一个小冰诗集。小冰说：“看那星，闪烁的几颗星，西山上的太阳，青蛙儿正在远远的浅水，她嫁给了人间许多的颜色”。**



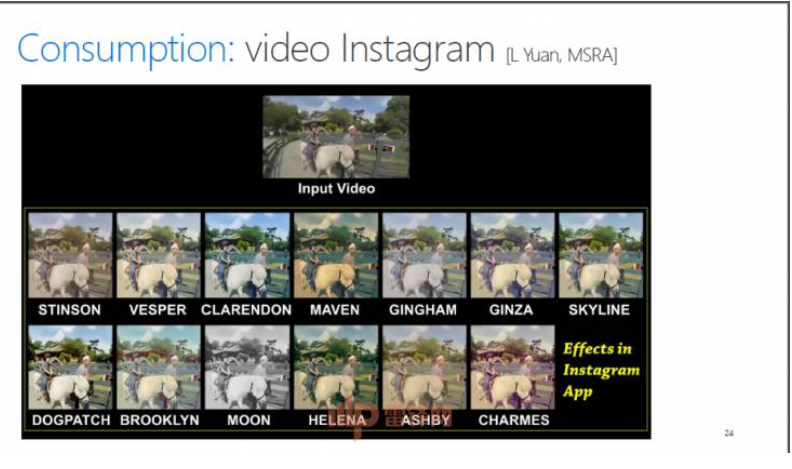
**我们另外还做style transfer（风格转移），给你一个油画或者卡通，能不能把这个风格转移到视频中，可以把这个水的波纹表达出来。**



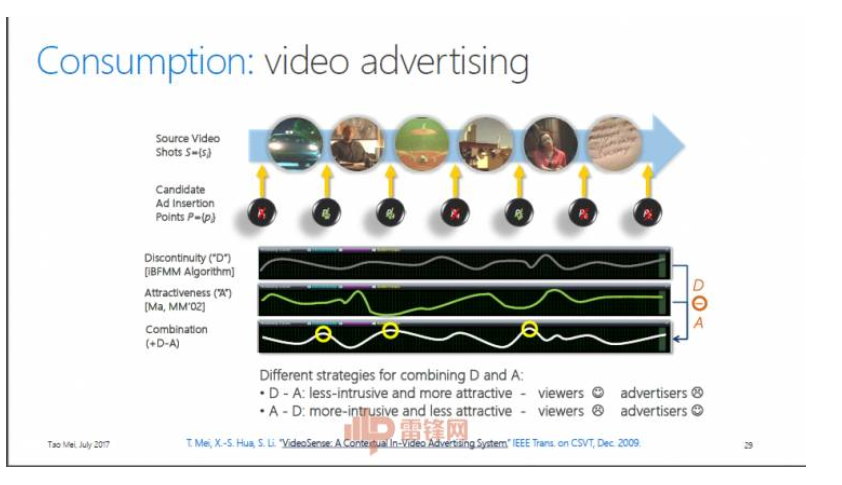
下面这幅图是某个娱乐节目，我们可以把这个人物分割出来放到另外一个虚拟的场景里面去。你可以想象，当两个人在异地谈恋爱的时候，我们能够给他一个房间，让他们在房间里面进行聊天。

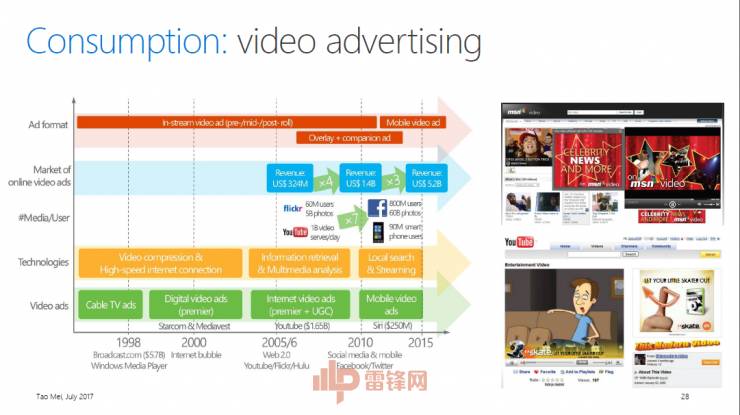


**还有Storytelling（讲故事），我能不能给你提供服务，让你的图片、视频变得更好，这都是机器产生的效果（见PPT）。这个风格叫Fashion，我们只要加上人工的处理，视频就可以让你的图片变得更加时尚。这很容易用在一些to C（针对消费者的市场）的场景里面。**



讲一下最后一个题目，这个广告是我十几年前加入微软的项目。那时候我们做的视频广告有两个问题需要解决：第一个问题是广告到底放在视频的什么位置；第二个问题是选什么样的广告，这个广告跟你插入点的信息是不是相关，使得用户接受度更加好。这两个问题怎么解决？当时我们提出一个方案，我来了一个视频，把这个视频分解，我们有几个值，第一个是discontinuity（不连续），看每一段是不是可以做广告，它的间断点使得用户的接受度更好。还有就是在是激动人心的阶段放广告。另外一个是Attractiveness（吸引力），让它变得可计算，当时我们有两个曲线，这两个曲线有不同的方式，第一种方式是要符合广告商的需求。

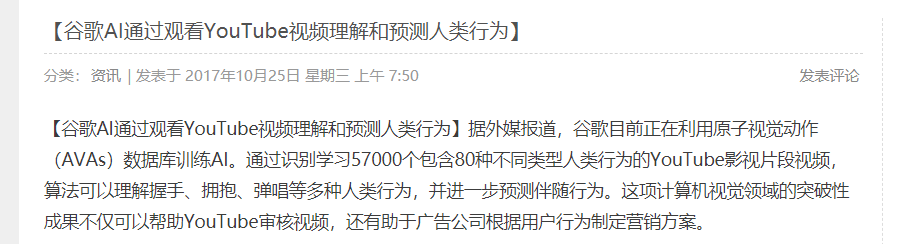


这个视频里面，当出现车子爆炸的镜头，我们可以识别出来这个内容，可以在这里放广告，使得广告和内容无缝连接在一起。我们也可以在故事需要的地方放广告。

刚才讲了很多场景和技术，但是在做科研的人看来，AI也好，深度学习也好，落地有很长的路要走，我们要脚踏实地的一个一个的去实现。

在1990年以前，学术业界开始有了一些关于视频合成技术和物体检测识别的讨论和研究。2000年以后，特别是2005年左右，整个学术业界开始泛起了对视频合成，视频内容理解的研究。据Google学术查阅的资料表明，在2005年以前，关于视频内容理解这一项就有4530条结果的收录。近十年来，这一学术搜索结果已增至17000条。同理，视频合成、物体检测识别、机器学习和NLP的理论研究也正如大家感受到的一样呈井喷态势。

**理解和预测人类行为：**



**Video Annotation：**

**行为分类 ( Action Recognition ) 和时序行为检测 ( Temporal Action Detection )：**

[**http://www.myzaker.com/article/59ba603a1bc8e0c05c000000/**](http://www.myzaker.com/article/59ba603a1bc8e0c05c000000/)

**学生总是把导师研究的东西当成科技前沿，研究的还是20年前的东西，这样一代传一代，形成一种阶级关系。**

[**https://github.com/sujiongming/awesome-video-understanding#video-classification**](https://github.com/sujiongming/awesome-video-understanding#video-classification)**（相关论文）**

