RACV2016 计算机视觉研究与应用创新论坛总结报告

李娜

2016.9.24

RACV2016 计算机视觉研究与应用创新论坛，该会议是我来到实验室第一次去参加的计算机视觉领域的会议，也是国内计算机视觉领域的主要学术活动。在上海科技大学由中国计算机学会(CCF)主办。会议持续三天，这三天也带给了我不同的经历。

9 月 18 日：

下午和俞老师还有小伙伴们去注册之后，去参观了博士生论坛墙报，刚开始进入会场其实是蒙圈的，墙报都是英文介绍的论文成果，自己只是对深度学习了解了 MLP 结构和 BP 算法，看到许多不认识的词汇以及不了解的专有名词，CNN、RNN 等名词不停的出现在墙报中，我的内心其实是崩溃的，很想去看懂一些什么，但是又听不懂他们在交流的内容。深感自己完全是一个“门外汉”，在跟随老师围着墙报转了一圈之后，决心要加强英语的学习，不产生畏惧心理，也要更深层地去学习 Deep learning 的知识（当时感觉自己不懂这个方向最基础的知识，也不好意思去问），还没有达到入门级别的自己，就好比一滴水遇到了一片海洋，渺小无知但又希望融入。

9 月 19 日：

早上有四位讲者作报告，首先是 Lihi Zelnik-Manor 讲述的在大量数据中分离出有用的可视化数据，英语不是很好，所以基本没有听懂，只是隐约能听懂几个和计算机视觉有关系的关键词，这学期一定好好学英语，多记录专业方向的名词，好多高端科技都是国外的专家作报告，所以学好英语也是一条必经之路。

林海滨，讲的通俗易懂，虽然帮助不是很大，但是我觉得增加了我对计算机视觉的应用认识，他提出了走红的 Pokemon Go 游戏，以及 AR,VR 的产品原理以及算法技术体系，体会到了在真实世界的信息叠加和全虚拟世界的肆意操作，还有 SLAM（即时定位与建图 simultaneous localization and mapping）三维场景感知，使用单目 SLAM 场景，支持场景存储，保持稳定表现，提升运算速度等。他演讲准备了很多体验视频没有播放出来，感觉有些遗憾。

李玺，主题是人工智能驱动的视觉特征计算、学习及应用。他讲述了人工智

能学习理论，首先是 MANN——学会学习（Learning to learn），其中有长短时模型（LSTM）通过隐性的共享记忆结构，不完全的实现知识的存储；神经图灵机（NTM）引入带记忆的神经网络去模拟大脑皮质的长时记忆功能，实现用极少量新人物的观测数据进行快速学习；记忆增强神经网络（MANN）提出一种新读写更新策略——LRUS（Least Recently Used Access），它每次读写操作只选择空间或最近利用的存储位置，读写策略完全由信息内容决定；其次是终身学习（Life-long Learning）是一种能够存储学习过的任务知识，利用旧知识快速学习新任务的完整系统方案，凭借人物间的知识共享和知识库的知识积累，突破了学习过程在样本集和时间上的限制，为实现高效及高度智能化的系统提供可能。

午饭过后，我们去看了 AR,VR 眼镜以及空中手写文字等技术展示，AR 没来得及排队有点遗憾，VR 以前体验过，用它来玩游戏感觉还是很真实有趣的。空中手写文字是制定手势识别之后开始写字，在手指指尖位置定标记，记录指尖的行迹，写完之后给出备选，选定即可结束。

下午，我们在学术报告厅听的专题竞赛，第一场是视频图像分析竞赛，分为多类对象检测和车辆类别检测，行人检索和物体检索。有两个挑战：小对象检测和目标宽高比不一致。检测优化有迭代检测和边界加权修正，队伍演讲的时候有些困，听的迷迷糊糊。

仔细聆听了第三场，爱奇艺视频标注竞赛。背景是在人工智能+的大数据时代，继续高效的智能应用将上传的视频进行标签分类，它的过程分为拆分、编目、审核、发布，竞赛需要通过计算机算法对输入视频进行自动标签识别，输出该视频所属标签，对海量视频自动化标注，进行多类分类(Multiclass Classification)和多标签分类(Multiabel Classification)，数据集有 20 个类别，每个标签有 2000 个训练视频和 500 个验证视频和 500 个测试视频。经过四个队伍的讲解，视频有三个特征，视频静态特征是关键帧和视频帧、动态特征是光流、音频特征是频谱图，对于每类特征，采用单独的 CNN 网络训练，3 个网络训练好后，对输入视频分别提取特征，特征融合，分类。特征融合采用 average pooling 把局部帧级别的特征聚合为全局视频级别特征，融合方式有 early fusion 和 late fusion，多分类问题采用 one-vs-all 策略，分解为多个二分类问题。技术框架见图 1，结论有对于单个模态，图像通道准确率最高；音频、光流对视频分类也有效；融合音频、光流可

以提高准确率，但提升较小。多标签策略，划分视频片段，每个片段采用单标签标注，整个视频，基于预测的类别概率，输出多个标签。有的队伍使用的深度 CNN，2CNN，里面团队还用到了深度残差网络（residual network 深度神经网络），有的团队在 flow 通道和 audio 通道同时使用残差网络，使得标注视频的正确率可以一步步提高。

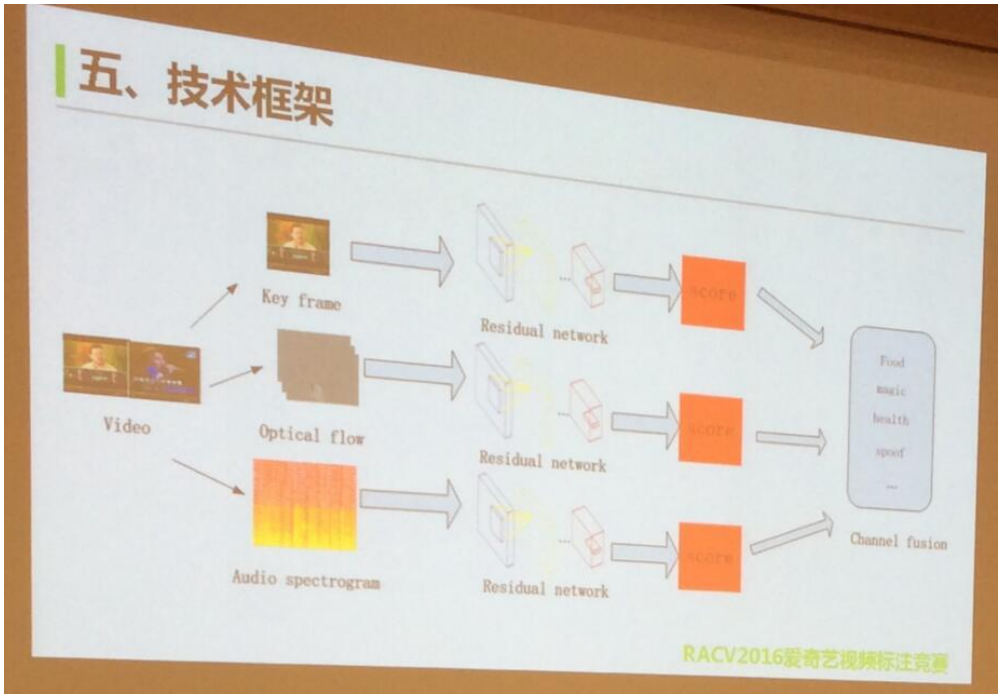

图 1. 视频标注技术框架

9 月 20 日：

权龙教授演讲的题目是：计算机三维视觉的发展与现状，向我们从最基础的讲述了什么是计算机视觉：像素、特征、相机、几何、分割、物体识别，从准密集方法、从点云到物体、街景三维重建、六点算法等，现在只需要一台无人飞行器就可以进行三维重建的图片采集，三维重建有许多行业应用，比如测量和绘制地图、建筑设计和遗址恢复、施工监督、文物保护及展示等等，很多专有名词不是很明白，但是知道我们计算机视觉实验室是和这些息息相关的，记下来以后总归再次看到学习的时候会记忆深刻。

陶海讲的计算机视觉应用，他讲述公司刚开始做图像识别，卖算法，做硬件，从端到端，业务需要定制化，进行有些艰难，图像识别处理，如果想要分辨率提升，计算能力也要提升，通过算法和硬件设备满足车牌识别、商场零售的客流估计和人机检测、景区限流等等。中间会遇到摄像机角度问题以及阴影成像问题，

他讲述说，能用器件解决的不用算法，而且不单单是技术的掌控，在艺术方面同样更胜一筹。感觉这个公司应该很有钱。能达到需求者需要的目的，看来钱不是问题，技术才是真正的核心。

吴毅红讲者讲述了基于图像的定位，讲了什么是相机定位：根据输入的 2D 图像或视频，计算出相机在三位物理空间中的位置和姿态称为相机定位，关键任务有匹配、相机标定、位姿计算，用到矩阵相乘的公式已经世界坐标系和摄像机坐标系，基于环境已知定位——基于点的通常称为 PnP；基于环境未知的定位，分为在线建模——基于视频 SLAM（单目 SLAM 和多目 SLAM），不在线建模——基于离散图像，SFM 的中间过程。

全程会议有很多很多很多听不懂的专业术语以及难以理解的英文讲解，自己学到的知识可以说是有限的，但是参观了这些大神们的演讲和竞赛，开阔了自己的眼界，之所以成为大神不单单是掌握了高端前沿的技术知识，还有流利简洁明了的表达和快速调度知识回答观众问题的能力。CNN 是会议多次提到的名次，我打算多加了解 CNN 来巩固我对这次会议许多报告的理解。我觉得日常不应该只注重知识的学习，拿来应用才是最有用的，希望自己可以增加实践操作能力。非常感谢老师给我去参会的机会，一程下来收获很多，希望可以内化吸收，做到更好。