10 机器感知

中科院研究员山世光：从看脸到读心：基于视觉的情感感知技术

整理：智源社区 张鲁洋

中科院计算所研究员山世光的报告主题是《从看脸到读心：基于视觉的情感感知技术》。

在报告中，山世光提到："人脸识别已近尾声，世间一切尽在脸上，我们能够从脸上看到更多的内容，若干年后我们很可能从"看脸时代"进入到"读心时代"。读心对于我们打造有温度、有情商的 AI 是至关重要的。"

所谓读心，即通过人的外显语言或行为来推测、估计出人的内在心理和精神状态，从情感角度对一个人进行评估。使用机器解决"读心"问题，可以更好地预知危险、评估风险，帮助警务人员破案等。

山世光在报告中介绍了他所带领的团队在该领域的一系列研究成果。包括精神疾病辅助诊断技术、人脸表情识别、面部动作检测、自监督学习以及基于视频的遥测式生理信号的测量等。当然，现阶段让机器像人一样具备察言观色的能力，还有很长的路要走； 山世光认为，我们不能操之过急，而需要先立足于解决现存的问题，比如弱信号检测、数据匮乏等。

以下为山世光的演讲正文：

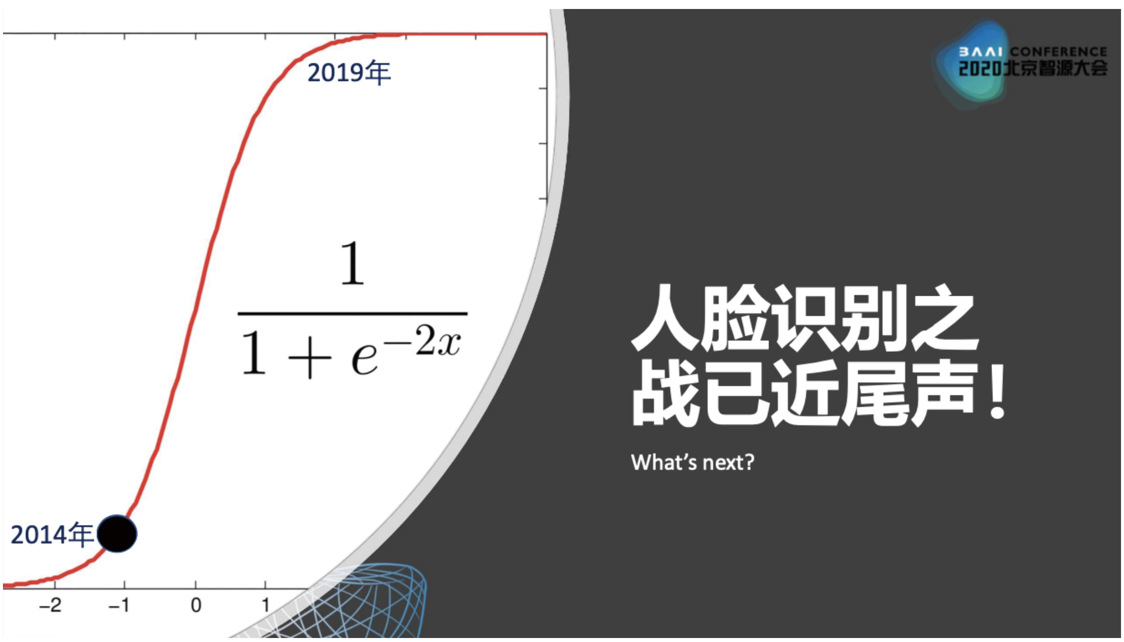


图 1： 人脸识别研究已接近尾声

在过去的八年中，人脸识别技术取得了非常大的进步。但从某种意义上说人脸识别研究已经接近尾声。但这并不意味着我们已经把人脸上的信息读完了——我们人类本身除了识别人脸，还能够从脸上得到更多的信息。例如中医可以望、闻、问、切，其中的望就是通过看脸来诊病； 普通人也都具备察言观色的能力，这是我们在人

类社会中生存的基本能力。因此，人脸技术的下一步也许将从“看脸时代”进入到“读心时代”。

读心，对我们打造有温度、有情商的 AI 至关重要。但什么是读心？本质上就是通过人的外显语言或行为推测、估计出人内在的心理和精神状态的技术，从情感角度对一个人进行评估。其内涵和外延有三个层次：

1) 瞬态特征，即生理指标，包括身高、体重、心律、呼吸、血压、血氧、眨眼率、视线等。
2) 短期特征，即心理状态，包括疲劳、专注、亢奋、无聊等内心在相对较短时间内的状态。
3) 长期特征，即精神状况，主要涉及到自闭症、抑郁症、焦虑、狂躁等，甚至是人格特质。

AI 如果能够识别出以上三个层次的心理和精神状态，在实际应用中将起到重要作用。比如，在自动驾驶领域，可以通过对司机疲劳状态进行评估，从而可以预防危险的驾驶行为； 在精神健康领域，可以通过计算机视觉技术，获取更多客观化的评估； 在机器人领域，可以利用这些技术让机器人拥有对人类情感进行理解和反馈的能力，使其能够更好地陪护人类； 也可以将这些技术应用于用户画像的评估，比如贷款风险评估、人力资源评估等。

在精神性疾病的辅助诊断方面（包括孤独症、抑郁和焦虑等），国内、外已经有很多研究者在进行研究：

2018 年，来自澳大利亚几个大学的联合团队采用多模态融合的方法 [1]，将语言特征（Paralinguistic ）、头部姿态特征 (Head pose) 以及眼神特征（Eye gaze behaviors）相结合，对抑郁症与健康人群的分类。融合之后的分类精度达到 88%。

2018 年，郭国栋团队也利用深度学习进行了抑郁症诊断的研究 [2]，他们通过融合面部特征（Appearance）和动态特征（Dynamics）进行抑郁状态的评估。在 AVEC2013 数据集（包括 82 个人、150 段视频）上进行实验，结果达到 7.58（MAE）、9.82（RMSE）。虽然误差还不是很理想，但精度却非常得好。

李飞飞团队在 2018 年也做了基于 3D 人脸表情和语音的抑郁程度评估工作 [3]，其使用的是 Casual CNN 方法，最终获得了 83.3% Sensitivity 和 82.8% Specificity 的实验结果。

国外也有一些人格计算（Personality Computing）方面的工作，即通过一个人的照片或视频，对其进行第一印象的分类。今年 6 月，Science Report 上发表了一篇来自俄罗斯团队的文章 [4]，他们通过照片评估一个人的 Big-five 人格分类 (外向性、亲和性、尽责性、神经质性和开放性)。这类工作在之前也有人研究，但是这篇文章使用了一个大规模的数据集，收集了 12447 人的 31367 张照片，每个人对自己的人格进行报告。他们用深度神经网络去做评估，结果显示模型的评估精度基本和人的直觉评估吻合。

事实上这项工作在国外很早就有研究，2014 年 IEEE Trans. on Affective Computing 就有一篇综述性文章 [5]，介绍了这方面的一百多篇文章，虽然采用的并不完全是视觉，但是不少工作都是基于视觉的方法来做的。

前面对从看脸到读心的变化做了概括性的介绍，下面介绍一下我们在学术上的一些工作，下面着重介绍几个方面。

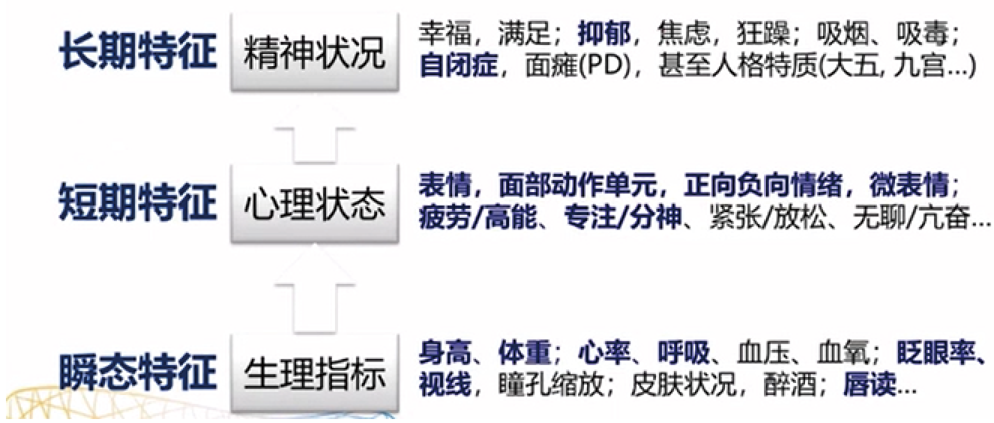


图 2：深刻理解人的视觉关键技术之内涵和外延

一、在自闭症儿童早期诊断上的应用探索

早期儿童抑郁症诊断的探索，首先是通过传感器希望记录儿童的社交行为或者个体行为的过程，然后利用计算机视觉的技术分析行为特征。为了让视线捕捉能够不依赖于专用的设备，目前我们希望能够采用普通的摄像头去做视线估计。ASD 或者孤独症儿童本身已经有非常多的工作，包括视点特征、表情特征、日常行为特征等，过去都有比较多的经验总结，我们依赖这些经验总结，重点去通过普通摄像头捕捉儿童的视线，然后分析他们的面部表情，以及在社交过程中和别人眼神对视以及互动的方式。

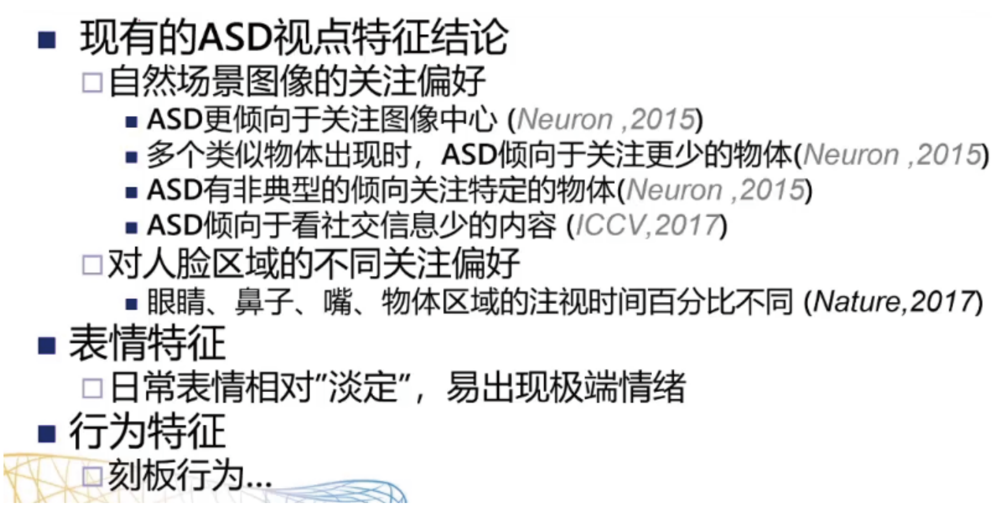


图 3：ASD 视点特征结论

目前涉及到的技术包括视线估计技术、表情识别技术、面部动作识别以及数据挖掘技术。我们也开展了这些方面的研究工作，例如在视线、眼神的估计方面，我们已经做了一些系统的工作；在专注度估计方面，2018 年我们也参加 EmotiW2018 挑战赛，通过视觉的方法估计一对一教学过程当中学员专注的等级。

在这一领域仍然面临着非常多的挑战，包括传感器和数据层面。特别是数据层面，目前我们发现几乎不太可能依赖有监督的、大规模的数据进行深度学习来实现这些功能，所以我们必须去研究如何能够在弱监督、无监督、

半监督、小样本等数据条件下，将知识嵌入到数据驱动中，使我们可以不依赖于大量的数据。

二、遮挡鲁棒的人脸表情识别

这个工作[6]主要是考虑人在做表情的时候经常下意识地用手遮挡自己的面部的问题。我们提出基于面部分块的门卷积神经网络，利用注意力机制自动感知人脸被遮挡的区域，进而增强非遮挡区域的重要性，最后结合局部注意力和全局注意力，使得我们能够保留对表情识别的全局和局部的信息。

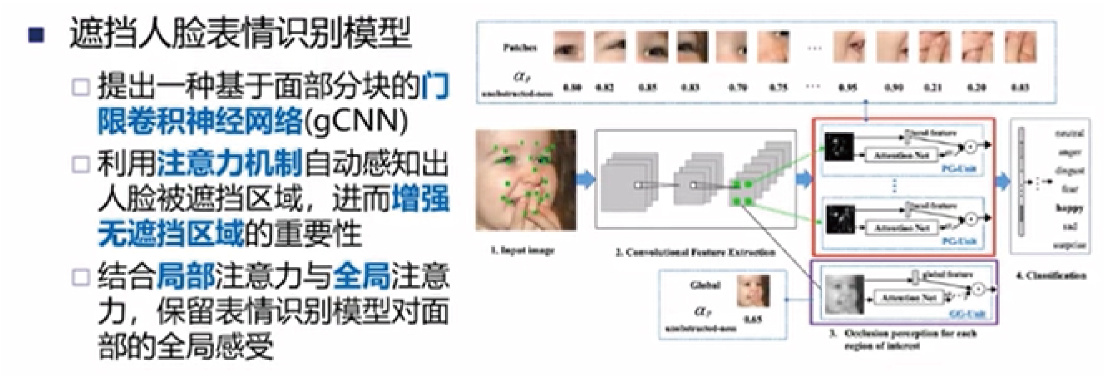


图 4：局部遮挡表情识别办法

我们根据人脸特征点对人脸分块，然后为每个区块学习分类“贡献”的权重。通过这样的机制，我们能够把更多的特征增强在非遮挡的面部区域，使得被遮挡情况下的面部表情识别结果更加鲁棒。我们的方法已经取得了明显优于之前算法的性能，在一些数据集上取得了比较优异的表现。

三、基于半监督学习的 AU 检测

刚才提到做表情识别、情绪识别等工作，但这些任务的数据集是不足的。那么我们有没有可能在有一部分标注数据和大量无标注数据的条件下去完成这些表情、情绪识别呢？去年我们在 NeurIPS 上发表的工作[7]就是试图解决这样的问题，即在半监督条件下如何做面部动作检测。

面部动作检测，我们又称之为 Action Unit，它是根据面部肌肉解剖定义出来的一些类似于眼睑下垂、闭眼、嘴唇收窄等动作，简称 AU。这个工作对数据标注是一个非常大的挑战，往往标注一分钟的视频可能需要一个专家花半小时以上的时间。这就导致这一领域的数据非常匮乏。那么有没有可能利用无监督的数据做这方面的工作呢？

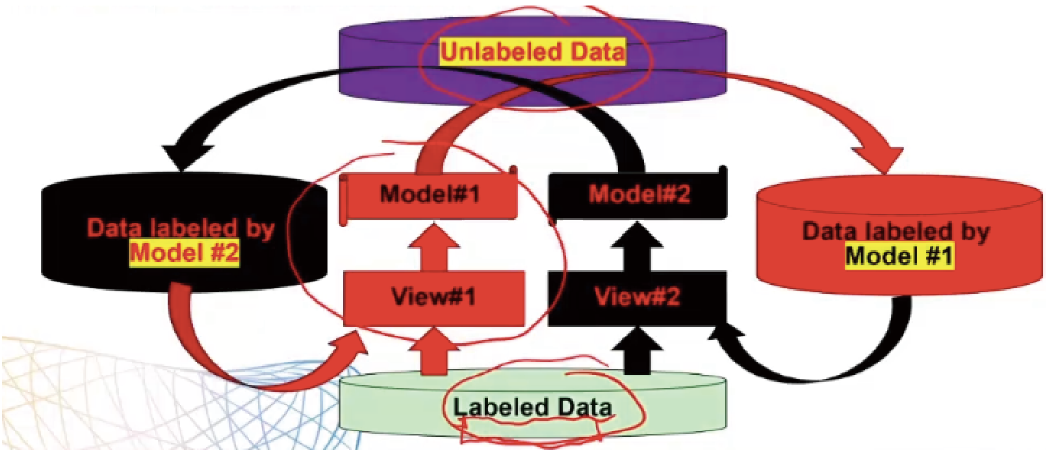


图 5：Co-Training 的方式同时利用 Label 和 Unlabel 的数据

我们采用 Co-Training 的方式同时利用 Label 和 Unlabel 的数据，学习两个模型。先用 Label 数据得到一个预测器，预测器会把所有 Unlabel 的数据进行 label 化，然后再和已有带有 ground truth 的数据结合，然后去训练第二个模型；第二个模型在把 unlabel 的数据 label 化，把这个结果和原有的 label 数据合并，然后再训练第一个模型。这是一个迭代的 Co-Training 过程。其实这种方式不是我们首先提出的，但我们采用了两个不同的 views，也就是两个不同的神经网络，两个网络会协同学习。

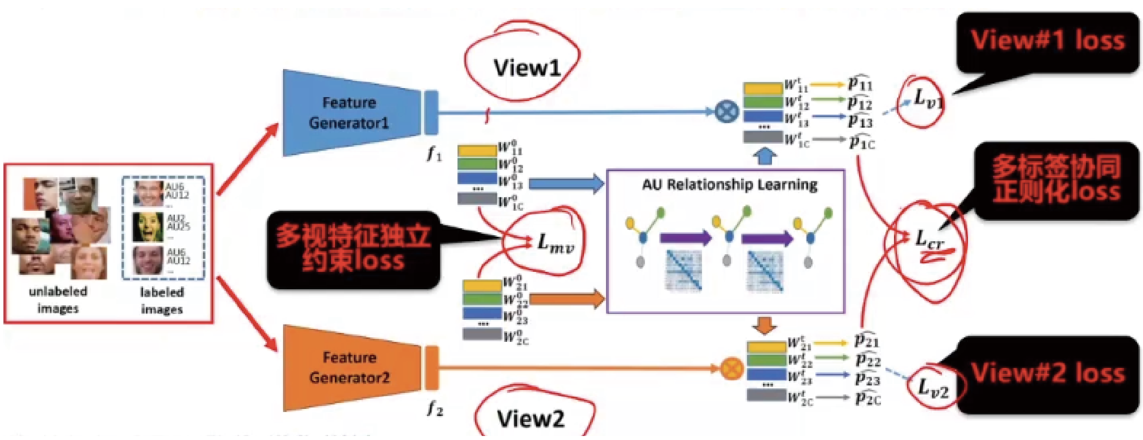


图 6：协同学习过程图

在协同学习的时候，我们为了保证两个网络有一定的独立性和相关性，会设计相应的 Loss 函数。同时我们也把多个不同的 views 之间进行协同，不同的面部动作单元之间也要协同，把不同的 loss 加进去作为目标函数，使得我们学习到两个不同的网络。这样的方法在 EmotionNet 数据集上的结果比之前高了 2.6%；在 BP4D 数据集上，比原来的方法高了 1.8%。

四、基于自监督表示学习的 AU 建模与检测

前面的工作是有一部分数据是监督的，另一部分数据是没有监督的。但我们觉得这还不够，我们希望做一个完全采用自监督方法学习 AU 的模型，这样才能够让我们在检测 AU 过程中大大减少对数据的依赖。所以这项工

作[8]的出发点就是希望利用大量带有情绪变化，但又无标注的视频数据，去更好地学习 AU 模型。

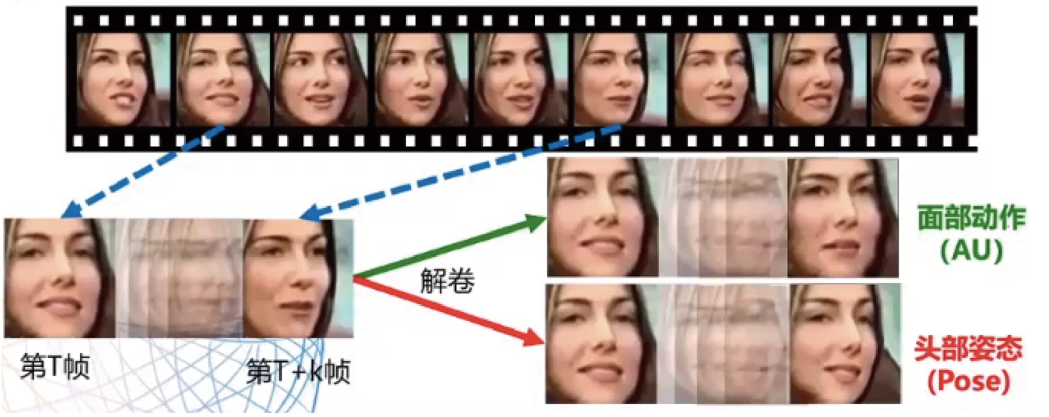


图 7：头部运动事实包含两部分

视频中头部的运动事实上包含两部分，一部分是面部动作导致的，另一部分则是头部姿态导致的。所以我们首先要做的是将这两部分动作进行解耦，然后保留面部变化的数据。

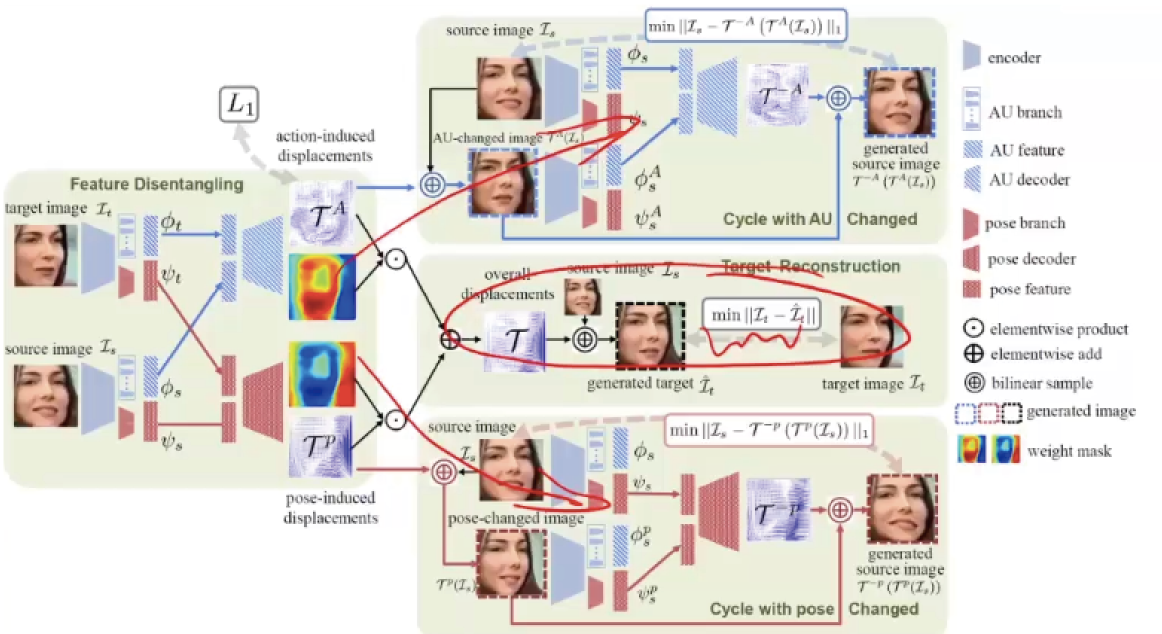


图 8：基于自监督方法学习 AU 的模型

为此我们设计了一个自监督模型，通过分解面部动作导致的运动和姿态导致的运动，然后再合成新的数据集。当我们分解的足够好时，就可以更好的从 T 帧去合成 T+K 帧。具体来说，我们采用了两个分支，一个是 AU，一个是姿态，最后通过合成去逼近 Minimum 误差。基于大量的训练，使得我们能够在 BP4D 数据集上取得和有监督的方法媲美的精度，甚至在 GFT 数据上我们也获得了比有监督方法更好的性能。

五、基于远距离普通摄像头的心率估计

上面是我们在面部表情方面的一些工作，特别是后面的两个工作我们都是希望能够在数据不够多的情况下有更好的 AU 检测和表情识别的精度，接下来再介绍基于远距离普通摄像头在心率估计方面的工作。在这一方面，从 2018 年开始，我们陆续发表了一些文章。[9][10][11][12]

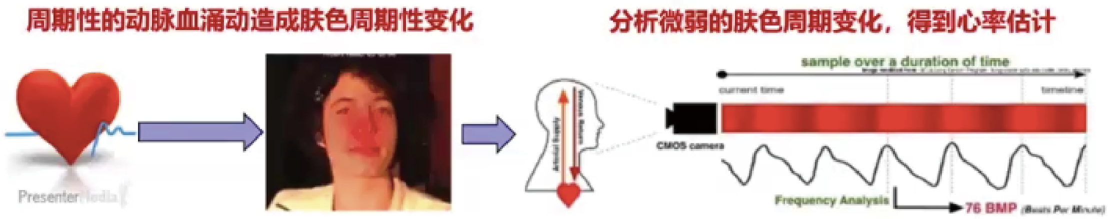


图 9：基于远距离普通摄像头的心率估计

我们人类不具有从远距离观看从而估计出一个人心率的能力，但是我们的心脏跳动会导致皮肤颜色出现周期性变化。医学上有关心率、呼吸率、血氧的测量，过去我们常用的是 PPG 技术，即根据血液对光的吸收强弱的原理，来实现对血液流量变化的检测。目前人们期望能够在远距离（例如 0.7 米）测量人的心率、呼吸率以及血氧等。

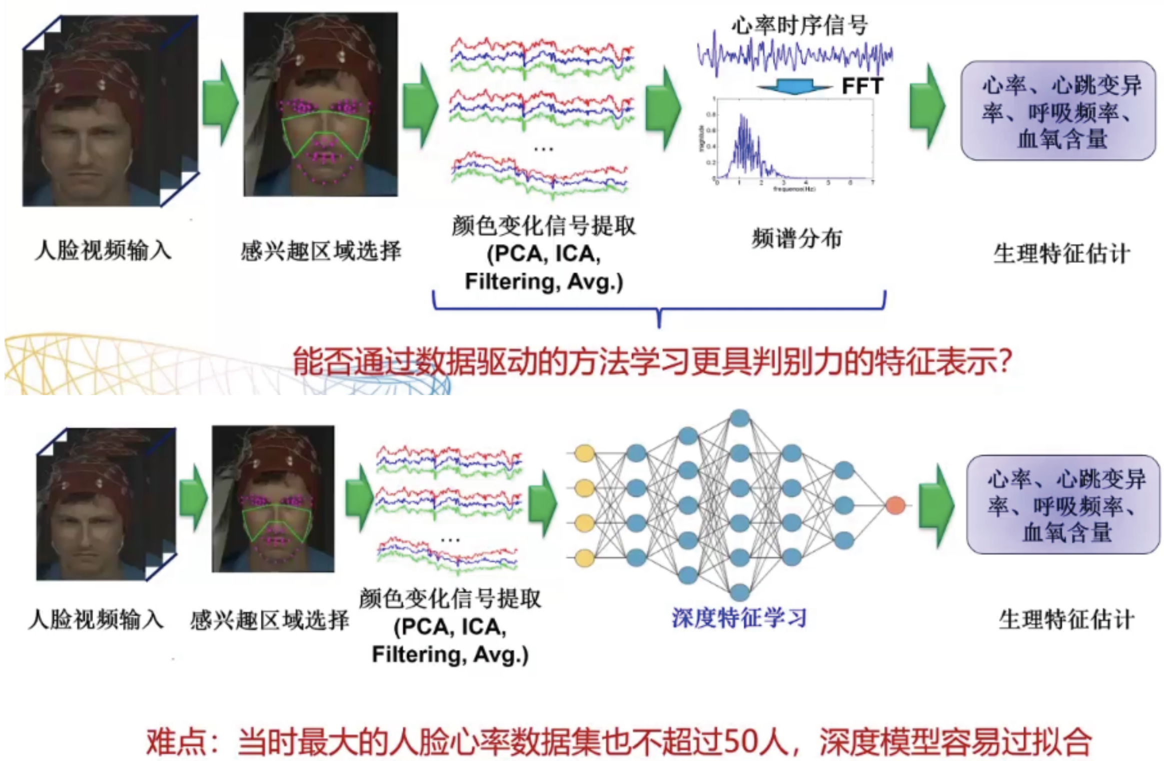


图 10：通过数据驱动的方法学习更具判别力的特征表示

通过对人脸拍摄视频来估计心率，这些工作从 2008 年开始引起大家的关注。之前的方法大都是采用对颜色变化进行独立的 PCA、ICA 以及 Filtering 等分析，从而获取特定的频率。那么，我们能否通过数据驱动的方法来

学习呢？当我们尝试去做时，发现结果并不好，因为这方面的数据非常少。比如在 2018 年时，最大的数据集不到 50 个人，很容易过拟合。

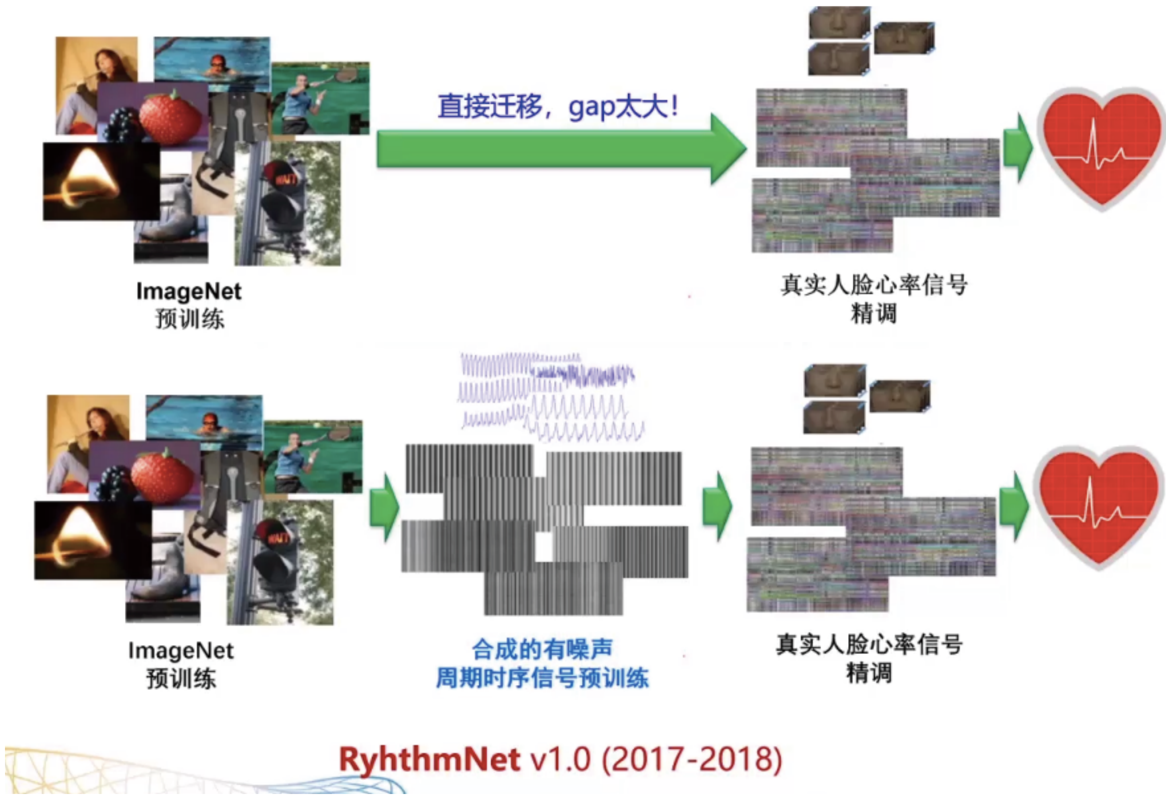


图 11：使用 RyhthmNet 做训练

一个容易想到的方法是使用迁移学习，但我们发现简单的迁移学习 gap 非常大。于是我们想到了一个新的方法[9]，即先合成一些周期性的时序信号，用这些合成的信号对模型做预训练，然后再用真实数据进行精细训练。我们发现这样可以获得不错的结果，在 MAHNOB-HCI 数据集上将 HRrmse 从过去最好的 6.23% 降到了 4.49%。

当然数据匮乏是一个很大的问题，因此我们自己也做了一个包含 107 人的数据集，包括了 3000 多段视频，涉及到不同光照、不同摄像头等。最近我们又发布了 500 人的数据。[10]

由于数据分布不均衡，大多数的心跳都分布在 60-90 这样的区间，这就导致我们很难估计其他区间的心跳。为了解决这个问题，我们采取了数据增广的方法，通过对视频的上下采样来模拟不同心率，通过这种扰动的方式我们可以获得更好的精度。为了解决头部运动干扰的问题，我们也有引入注意力机制，在我们自己的数据集 VIPL-HR 上将 RMSE 提高到了 7.99。[11]

为了更好地用深度学习方法进行学习，我们将视频数据扩展成二维时空，然后直接用 CNN 对其进行估计，在时序上利用 RNN 建立相邻视频片段关系。最后在 MAHNOB-HCI 数据集上将 RMSE 提升到了 3.99，在我们自己的 VIPL-HR 数据集上得到了 5.3 的 MAE。[12]

六、唇语识别在路上

最后介绍一下我们在唇语识别方面的工作。我们发布了 LRW–1000 数据集，数据集包含了 500 多小时的原始视频，1000 个汉语词。

图 12：LRW 数据集

另外，在牛津大学发布的 LRW 数据集上我们获得了 87.3% 的准确率。在我们自己发布的数据集中，最开始的精度是 38.19%，现在已经提升到了 56.85%。在句子级唇语识别上我们获得了 11.2% 的错误率。在私有指令级比如“打开后车窗”、“打开导航”等短语句子的识别上，准确率达到了 93%。

七、总结

在人脸识别得到非常好的解决之后，我们认为未来会有更多的希望利用人脸技术去理解人的情感。当然，这项工作非常有挑战性，距离我最开始介绍的“像人一样察言观色”的能力还有很长的路要走。但是在特定的应用上，我们已经可以做很多事情，比如疲劳检测，现在已经逐渐落地；当然即使这些已经成熟的技术也有很多挑战，以呼吸率、心率和血氧估计的任务为例，它们面临着严重的弱信号检测问题，如何提高信噪比是值得关注的。最重要的是方法层面，因为数据的匮乏，我们必须要去研究如何把领域知识、医学知识、健康知识引入到弱监督、小样本数据这样的机器学习中，使我们能够更好的解决这些问题。因此，可以说人脸表情识别任重而道远，仍然有很多值得学术领域去做的基本问题，希望有更多的老师和同学一起合作来开展这些研究。

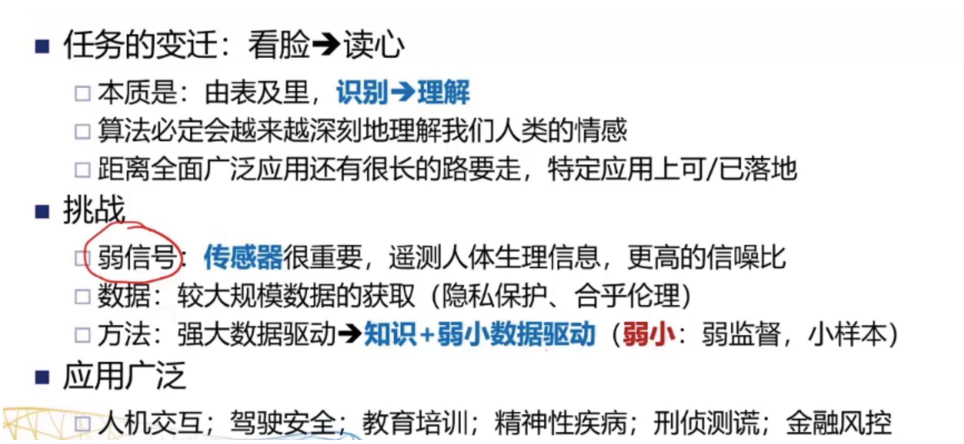


图 13：人脸识别的挑战和具体应用

参考资料：

[1] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors. IEEE T on Affective Computing. 10-12 2018;

[2] Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics. IEEE T on Affective Computing, 2018

[3] Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions. Machine Learning for Health (ML4H) Workshop at NeurIPS 2018. By Li Fei-fei group

[4] Kachur, A., Osin, E., Davydov, D. et al. Assessing the Big Five personality traits using real-life static facial images. Sci Rep 10, 8487 (2020). https://doi.org/10.1038/s41598-020-65358-6

[5] A. Vinciarelli and G. Mohammadi, “A Survey of Personality Computing,” in IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 273-291, 1 July-Sept. 2014, doi: 10.1109/TAFFC.2014.2330816.

[6] Yong Li, Jiabei Zeng, Shiguang Shan and Xilin Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, IEEE Transactions on Image Processing. 28(5), pp2439-2450, 2019

[7] Xuesong Niu, Hu Han, Shiguang Shan, Xilin Chen. Multi-label Co-regularization for Semi-supervised Facial Action Unit Recognition. NeurIPS 2019

[8] Yong Li, Jiabei Zeng, Shiguang Shan, Xilin Chen. Twin-Cycle Autoencoder: Self-supervised Representation Learning from Entangled Movement for Facial Action Unit Detection. IEEE/CVF CVPR2019

[9] Xuesong Niu,Shiguang Shan,Hu Han,and Xilin Chen.SynRhythm:Learning a Deep Heart Rate Estimator from General to Specific.ICPR 2018.

[10] X.Niu,H.Han.S.Shan,and X.Chen.VIPL-HR:A Multi-modal Database for Pulse Estimation from Less-constrained Face Video.ACCV 2018.

[11] X.Niu,X.Zhao,H.Han,A.Das,A.Dantcheva,S.Shan,and X.Chen.Robust Remote Heart Rate Estimation from Face Utilizing Spatial-temporal Attention.IEEE FG2019(最佳海报论文)

[12] Xuesong Niu,Shiguang Shan*,Hu Han,Xilin Chen.RhythmNet:End-to-end Heart Rate Estimation from Face via Spatial-temporal Representation.IEEE Transactions on Image Processing.2020

中科院研究员王亮：面向复杂任务的视觉认知计算

整理：智源社区 沈磊贤

王亮本次的报告主题是《面向复杂任务的视觉认知计算》。

王亮，中国科学院自动化研究所研究员，模式识别国家重点实验室副主任，中国计算机学会计算机视觉专委会副主任。曾获得国家杰出青年科学基金、中国青年科技奖，是 IEEE TPAMI、IEEE TIP、Pattern Recognition 等国际知名期刊的编委。主要从事计算机视觉、模式识别、机器学习等相关领域的研究。

在报告中王亮介绍了基于深度学习的视觉认知机制建模，包括注意、记忆、推理、反馈等，并用于提升视觉语言匹配、视觉语言描述等复杂视觉任务的性能。

一、研究背景和挑战

1.1 研究背景

基本的视觉任务，比如目标检测、目标分割、目标识别等，在视觉认知计算领域已经得到了广泛的研究。但是复杂的视觉探索相对较少，这里的复杂视觉任务指的是多模态的学习任务，这样的任务除了视觉之外，还与文本、语音等相关联。复杂视觉任务中几个具体的研究问题如图 1 所示，比如视觉问答（给定图片，描述图中内容）、跨模态检索（给定图像检索语义相关段落，或者通过文字检索图片）等。

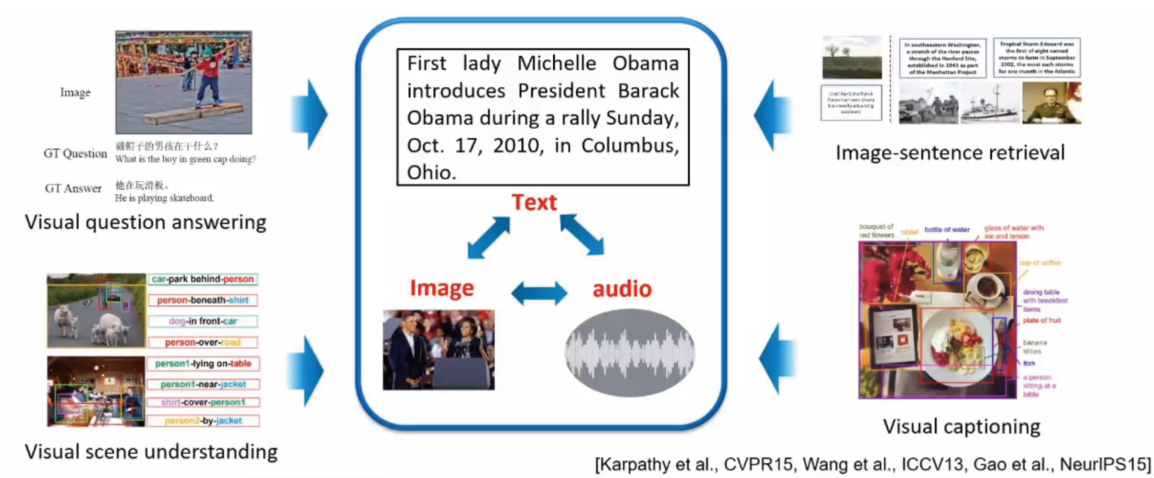


图 1： 复杂视觉任务的研究问题

复杂视觉任务有很多潜在的应用，但相比较于基本的视觉任务，复杂视觉任务研究起来也面临更多的挑战，因为不仅要处理视觉任务中的高级语义问题（譬如关系的推理、运动的推理等），同时还要处理与文本、语音等跨模态的交互、关联问题。

1.2 研究挑战

以视觉与语言（Vision and Language）场景为例，复杂视觉任务研究中存在 4 个挑战性问题：

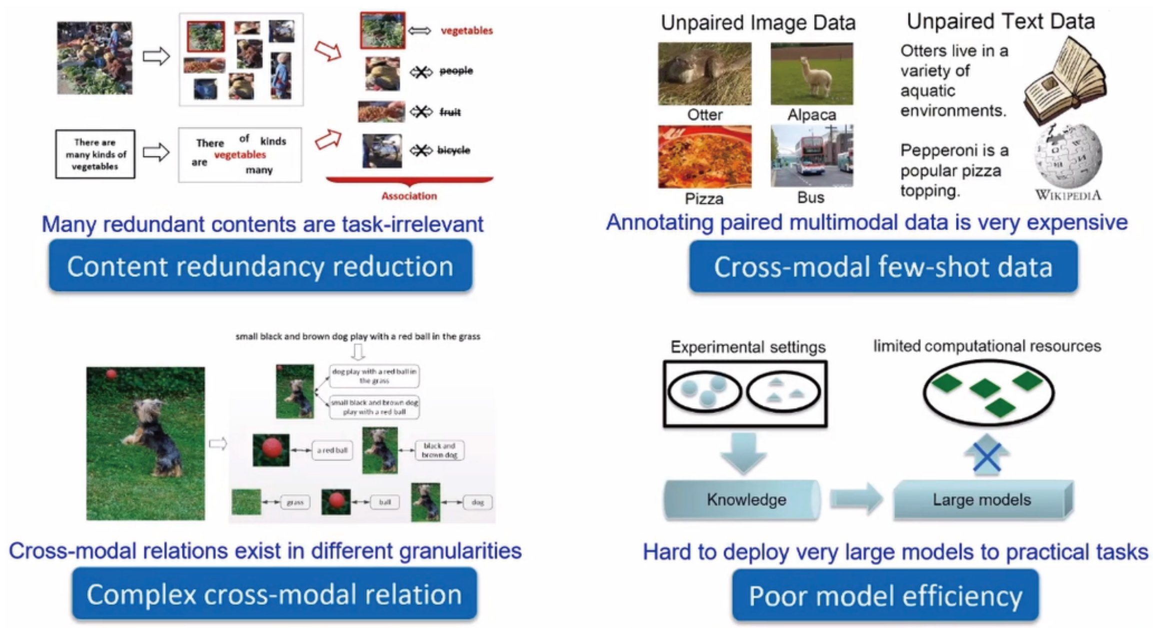


图 2：复杂视觉任务的四个挑战性问题

数据内容的冗余：以上图左上角的图片和文字段落为例，若衡量这二者之间的语义相关性，可以发现，图像中表示“蔬菜”的区域和句子中“vegetable”（蔬菜）这个单词存在语义相关性，但对于其他的单词或者图像区域没有直接的关联，所以类似这样的剩余信息就是一种与任务无关的干扰信息，且难以去除。

小样本问题：在多模态的情况下，样本分布存在小样本问题，要标注成对的模态是非常耗时耗力的。

模态之间的异质性：在视觉 – 语言的场景下，存在复杂的视觉语义鸿沟问题，图像和文本从局部到整体的不同层面可能存在复杂的对应关系，解决这种复杂的跨模态关系是一个重要的挑战。

模型的泛化性能差：模型在特定条件下表现良好，而在不同环境中性能则大打折扣。此外，很多相对高精度的工作是依靠复杂的模型堆叠处理的，现实的场景下很难进行高效率的部署。

总之，在数据理解时牵涉到一些认知的功能，比如信息的过滤、存储、再使用或者信息的推理等，这些功能并不能通过当前简单的视觉感知的计算来实现。

1.3 相关研究进展

在研究进展方面，王亮在报告中介绍了国内外在视觉感知和认知领域的近况。

视觉感知计算，如图 3 中左侧所示，通过借鉴生物学神经网络的结构，来感知视觉信息当中的形状、色彩和运动以及相关的信息。王亮认为在感知层面，很难处理复杂视觉任务当中的信息冗余以及复杂的关系推理。他举

出了视觉感知计算在视觉问答、视觉对话、视觉描述和跨视觉检索等任务上的表现，如图 3 右侧所示，即便是目前最好的模型，在约束受限的数据库上进行测试，准确率 (2019 年) 只有 70% 左右，远低于人的推理感知能力。

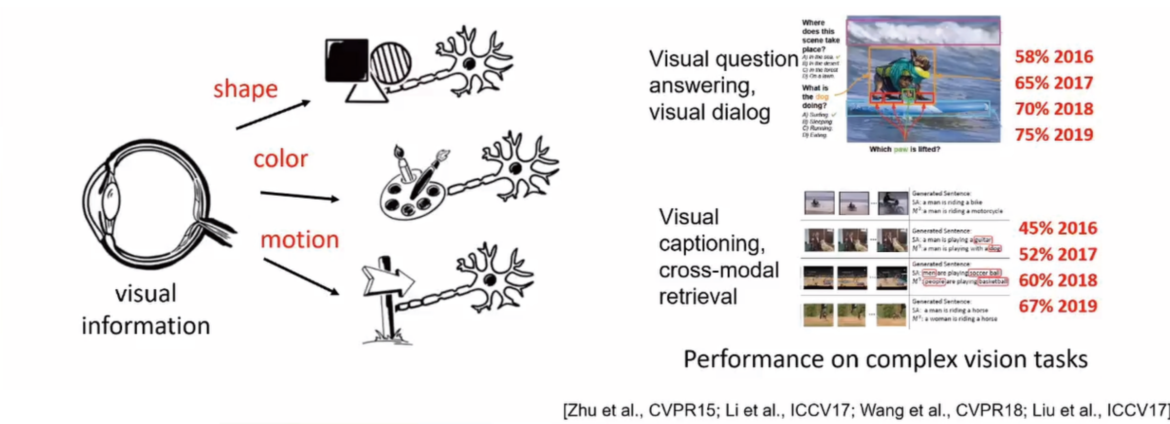


图 3：视觉感知计算

因此，研究者想在感知计算的基础上引入认知计算，来解决这样的问题。目前计算机视觉领域主要在探索视觉注意力机制和记忆机制，从而可以实现信息的滤波、存储、比对和推理等认知功能。王亮认为，将这样的研究思路借鉴到复杂的视觉任务当中，就可以实现一些小样本分析、知识的 Transfer，关系的推理以及决策等复杂的问题。

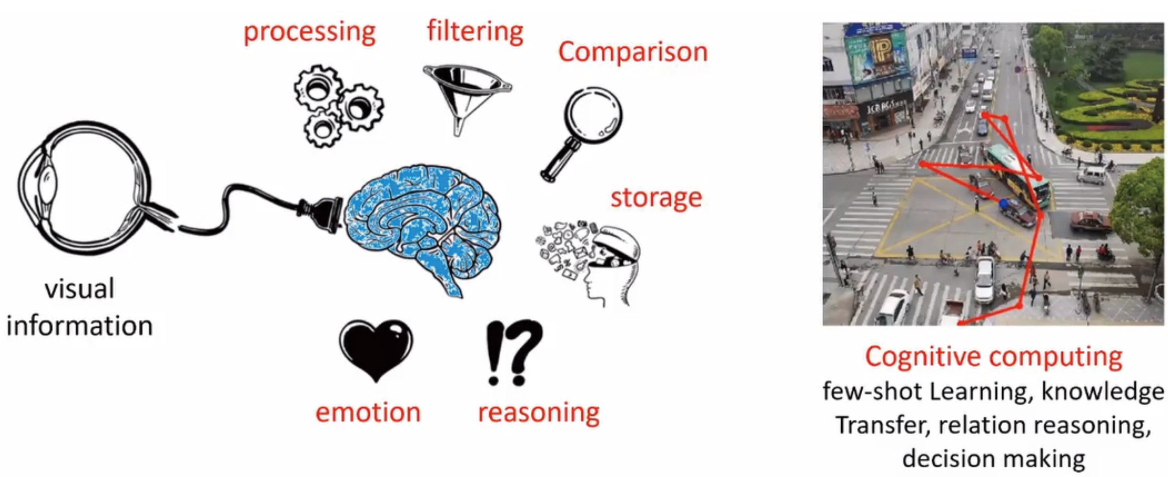
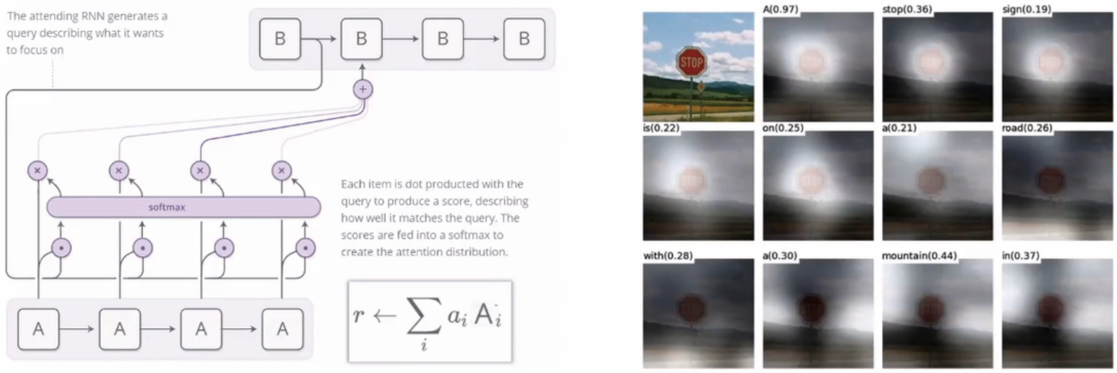


图 4：在感知计算的基础上引入认知计算

目前的相关研究进展，主要有注意力机制、记忆机制和推理机制三个方面的建模。

注意力机制建模。主要分为软注意力机制 (Soft Attention) 和硬注意力机制 (Hard Attention)。图 5 左侧给出了软注意力机制计算的模式，对不同的局部特征分别预测其权重值。注意力机制建模后的结果是一个加权和，这样权值大的局部特征将主导注意后的特征。硬注意力机制不采用加权和的方式，而是选择权值最大的局部特

征作为整体特征。在实际应用中 soft attention 效果要更好一些，因为 Hard Attention 会丢失很多的信息。



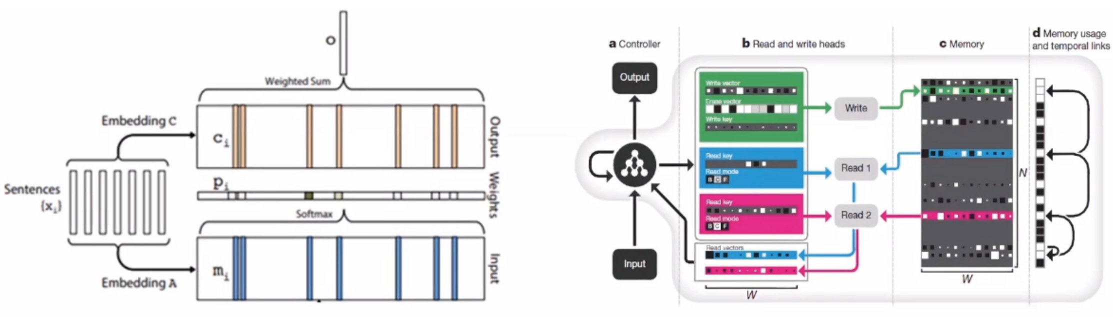
Details of soft attention

Attention maps in image captioning

[https://distill.pub/2016/augmented-rnns/, Xu et al., ICML15]

图 5：注意力机制建模

记忆机制建模。主要两种形式，图 6 左侧中的形式是端到端的记忆模式，代表短时记忆建模，通常具有读取的功能，但不具备写入的功能，记忆特征的初始化选择已有的样本等。图 6 右侧为神经图灵记忆，可以对长时记忆进行建模，具有读取和写入功能。这中机制是更一般的形式，记忆特征的初始化一般是随机特征进行开始，目前记忆机制的建模在序列化的预测任务中展示了较强的作用，用于建模长时间的相互依赖关系。



End-to-end memory
(memory items can only be addressed)

Neural Turing memory
(memory items can be updated)

[Sukhbaatar et al., NeurIPS15; Graves et al., Nature16]

图 6：记忆机制建模

推理机制建模。主要是推理不同的视觉目标、属性以及行为之间的关联关系。图 7 左侧是一个机器人导航的例子，从出发点进行推理决策，到目标点的过程，右侧是视觉关系的推理，比如空间关系，「天空在树之上」，「树在天空之下」这种视觉目标关系的推理过程。在推理机制建模中使用比较多的是强化学习和图卷积神经网络，因为两者可以很好地建模数据之间地关系，而且推理常常不是一次完成的，需要反复循环迭代得到最终的结果。

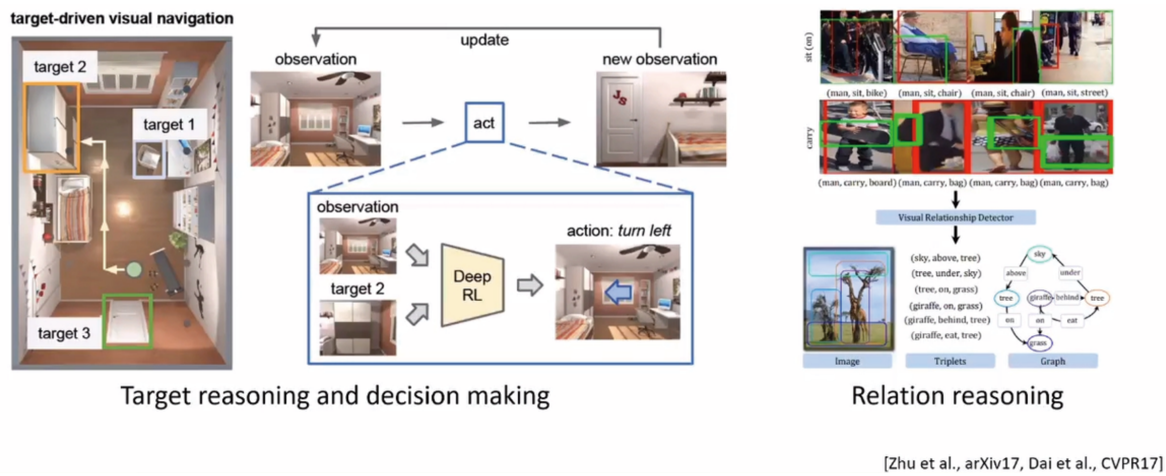


图 7：推理机制建模

二、团队工作

王亮介绍了他们团队的几个工作，也是从注意建模、记忆建模和推理建模三个方面展开。

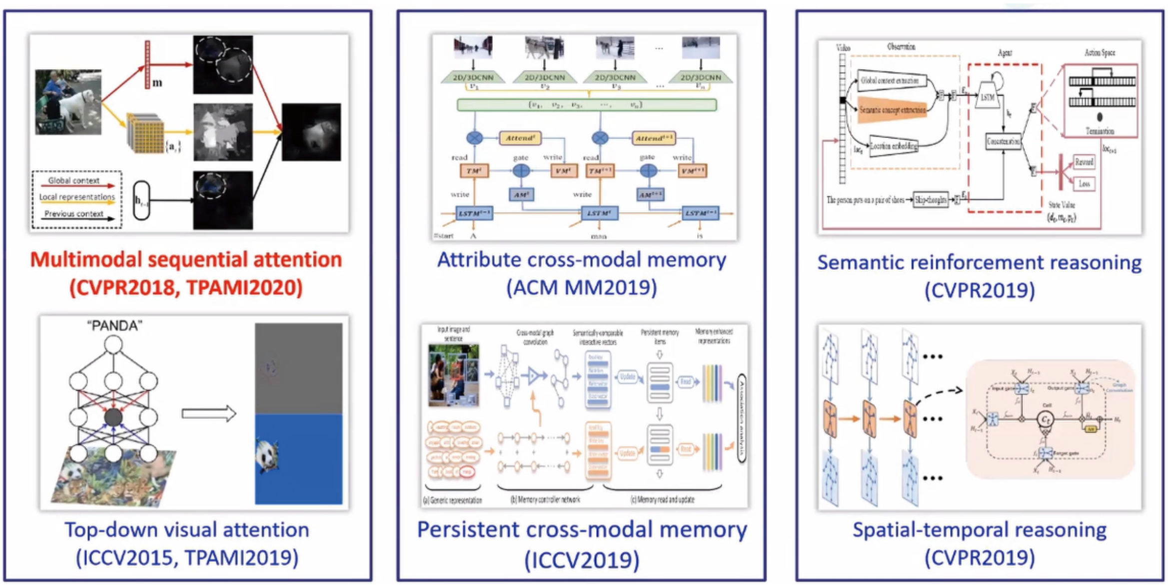


图 8：王亮团队在注意力建模、记忆建模、推理建模方面的工作

2.1　注意力机制建模

首先是注意建模，发表在 CVPR 2020 上。这项工作面向图像与句子的匹配任务。顾名思义，图像与句子的匹配其实是描述二者之间的相似度，应用场景十分广泛，比如图像句子的跨模态检索、图像描述、图像问答等等。

在图像与句子的匹配方面，传统研究方法提取图像或者句子的全局特征，使用结构化的损失函数或具有正则关系的目标函数进行关联。但在实际过程当中，王亮团队发现，无论是句子还是图像除了语义相关的一小部分外，其它的都是与任务无关的背景噪声，如果直接使用全局的图像特征并不合适。

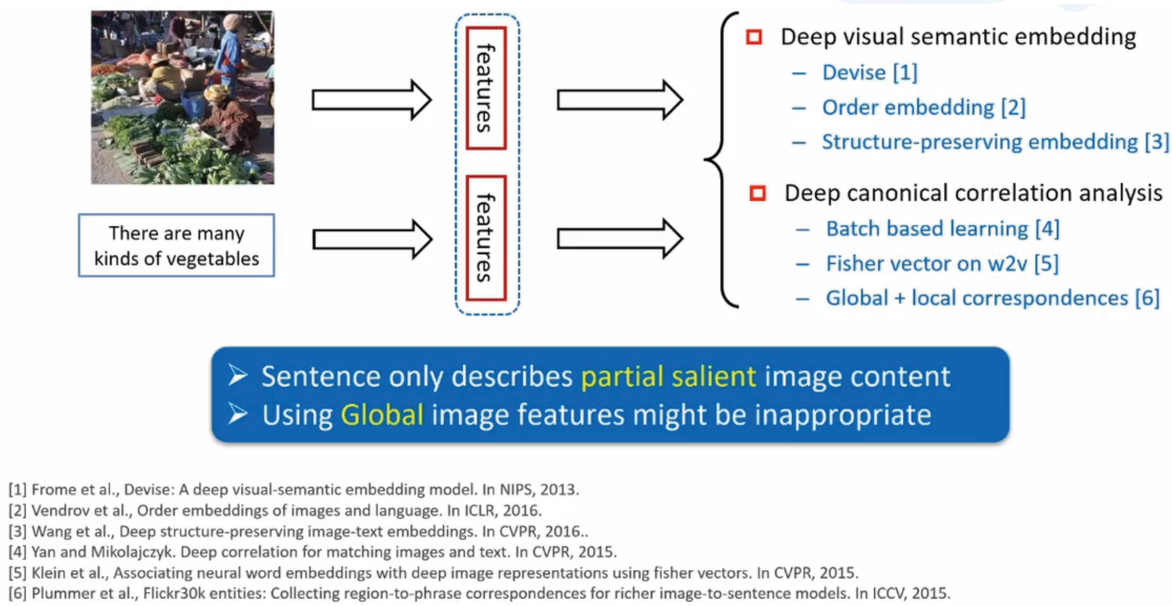


图 9：注意相关工作

王亮团队通过语义概念的提取和语义顺序的组织来解决上述问题： 采用使用多区域、多标签的卷积神经网络提取语义概念；使用上下文调制的策略学习语义顺序，使用句子生成作为指导。

王亮首先介绍了语义概念和语义顺序的定义。语义的概念即图像当中的目标、目标的属性以及目标行为，语义的顺序指的是语义概念之间组成一个句子的先后顺序。只有学习合适的语义概念，然后以一定的语义顺序组织以后，才能以一个比较准确的句子来描述图像中的内容。以下图为例，与之相匹配的语句是“一个快速奔跑的猎豹在草地上追逐小羚羊”，如果把“羚羊”和“猎豹”顺序颠倒的话，语义顺序就完全不一样。

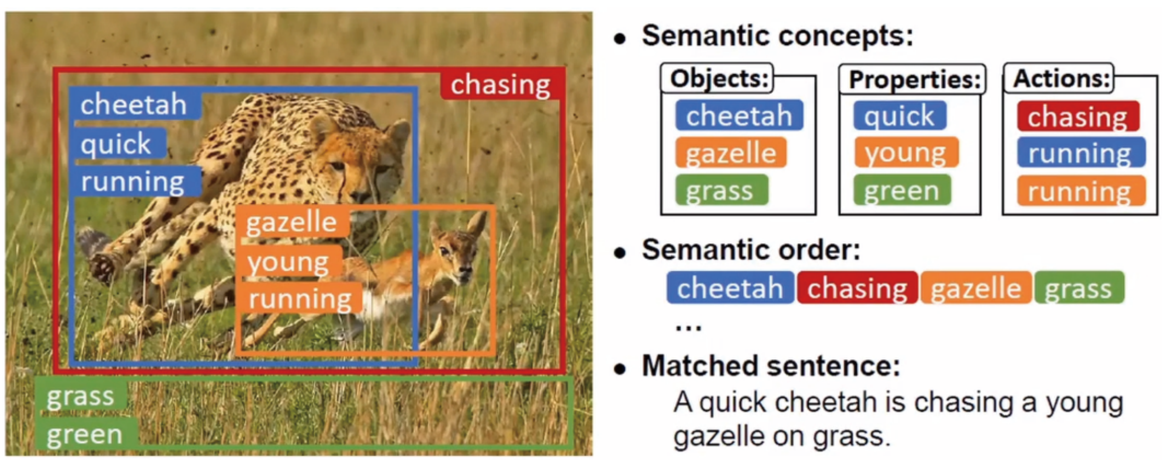


图 10：语义顺序

对于语义概念的提取，王亮团队采用是多区域、多标签的 CNN 进行实现，如图 11 所示，由于概念的生成没有现成的数据集进行处理，所以他们使用当前数据库当中的句子进行处理，选择理想的概念并且减少词汇表中词汇的数量，同时使用多标签、多区域的 CNN 进行概念的预测。

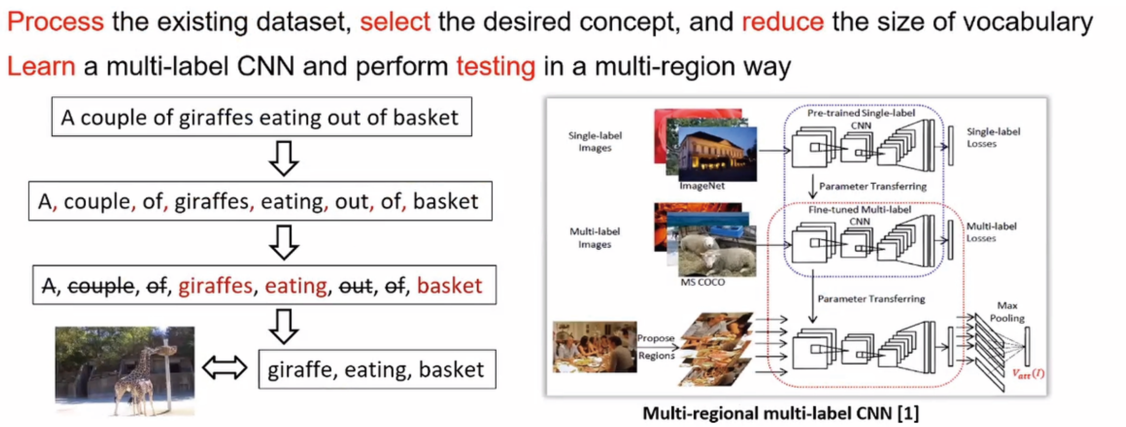


图 11：多区域、多标签的 CNN

对于语义顺序的学习，王亮团队使用全局的上下文特征作为参考。全局的上下文特征标注了语义概念在空间上的关系。选择性地平衡语义的概念和全局上下文之间的重要性，并利用句子生成作为一种指导，把融合后的全局上下文语义概念作为图像的表达，同时使用真实的语义顺序监督图像表达的学习过程。

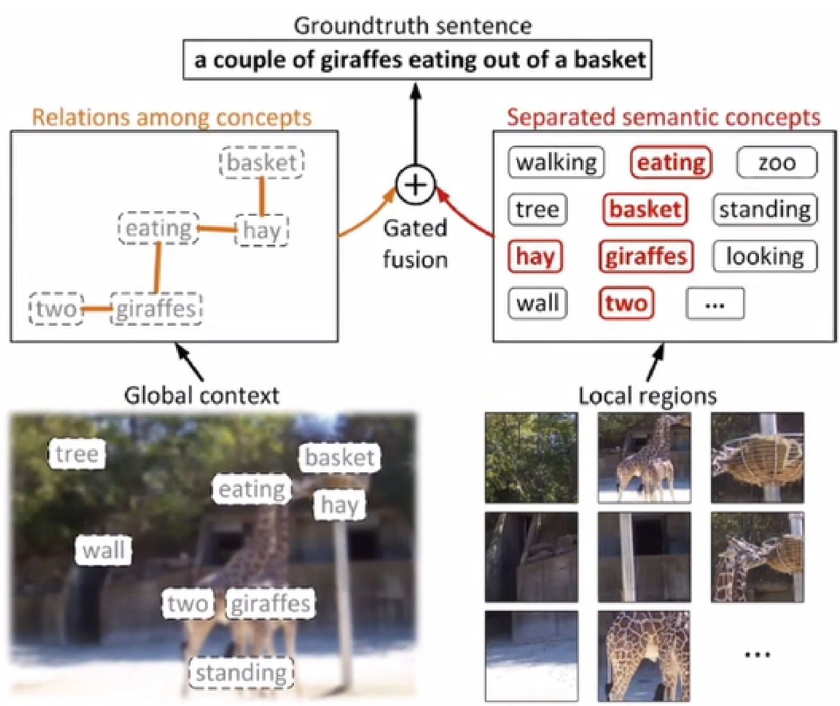


图 12：全局上下文特征标注

该模型在几个典型的数据库上进行实验。主要有两种任务:

1. 图像标注（image annotation）即给定图像，检索相匹配的句子;
2. 图像检索，给定一个句子，检索对应的图像。

如图 13 所示，实验结果表明上述方法在这两个数据集都获得了最好的性能。

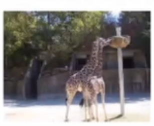
Bidirectional image and sentence retrieval results on two datasets.

	Method	Flickr30k dataset							MSCOCO dataset						
		Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VGGNet	m-RNN [21]	35.4	63.8	73.7	22.8	50.7	63.1	51.6	41.0	73.0	83.5	29.0	42.2	77.0	57.6
	FV [14]	35.0	62.0	73.8	25.0	52.7	66.0	52.4	39.4	67.9	80.9	25.1	59.8	76.6	58.3
	DVSA [12]	22.2	48.2	61.4	15.2	37.7	50.5	39.2	38.4	69.9	80.5	27.4	60.2	74.8	58.5
	MNLM [13]	23.0	50.7	62.9	16.8	42.0	56.5	42.0	43.4	75.7	85.8	31.0	66.7	79.9	63.8
	m-CNN [20]	33.6	64.1	74.9	26.2	56.3	69.6	54.1	42.8	73.1	84.1	32.6	68.6	82.8	64.0
	RNN+FV [16]	34.7	62.7	72.6	26.2	55.1	69.2	53.4	40.8	71.9	83.2	29.6	64.8	80.5	61.8
	OEM [29]	-	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	68.6
	VQA [18]	33.9	62.5	74.5	24.9	52.6	64.8	52.2	50.5	80.1	89.7	37.0	70.9	82.9	68.5
	RTP [26]	37.4	63.1	74.3	26.0	56.0	69.3	54.3	-	-	-	-	-	-	-
	DSPE [32]	40.3	68.9	79.9	29.7	60.1	72.1	58.5	50.1	79.7	89.2	39.6	75.2	86.9	70.1
	sm-LSTM [11]	42.5	71.9	81.5	30.2	60.4	72.3	59.8	53.2	83.1	91.5	40.7	75.8	87.4	72.0
	2WayNet [4]	**49.8**	67.5	-	**36.0**	55.6	-	-	55.8	75.2	-	39.7	63.3	-	-
	DAN [23]	41.4	73.5	82.5	31.8	61.7	72.5	60.6	-	-	-	-	-	-	-
	VSE++ [5]	41.3	69.0	77.9	31.4	59.7	71.2	58.4	57.2	85.1	93.3	45.9	78.9	89.1	74.6
	Ours	44.2	**74.1**	**83.6**	32.8	**64.3**	**74.9**	**62.3**	**66.6**	**91.8**	**96.6**	**55.5**	**86.6**	**93.8**	**81.8**
ResNet	RRF (Res) [19]	47.6	77.4	87.1	35.4	68.3	79.9	66.0	56.4	85.3	91.5	43.9	78.1	88.6	73.9
	DAN (Res) [23]	55.0	81.8	89.0	39.4	69.2	79.1	68.9	-	-	-	-	-	-	-
	VSE++ (Res) [5]	52.9	79.1	87.2	39.6	69.6	79.5	68.0	64.6	89.1	95.7	52.0	83.1	92.0	79.4
	Ours (Res)	**55.5**	**82.0**	**89.3**	**41.1**	**70.5**	**79.9**	**69.7**	**69.9**	**92.9**	**97.5**	**56.7**	**87.5**	**94.8**	**83.2**

图 13：实验结果

王亮进一步给出了一些具体的例子，如图 14 所示。真实的匹配句子是红色标记，用下划线标注的句子是共享相似语义。表格从左向右，第 2–4 列分别是上下文、概念 + 上下文、概念 + 上下文 + 句子生成，可以看出随着各种方法的加入，性能的指标越来越好。

Results of image annotation by 3 ablation models.

Query	Retrieved top-4 relevant sentences		
	ctx	cnp + ctx	cnp + ctx + gen
① 	1. a dinner table with various plates of food and a glass of water on the table 2. <u>a table top with some plates of food on it</u> 3. a table set for three with food and wine 4. a dinner table with three plates of gourmet hamburgers	1. <u>a meal is being displayed on a table</u> 2. **a table with bowls of grains and fruit and a hand with a plate** 3. <u>a table top with some plates of food on it</u> 4. **a person holding a bowl of oats next to bowls of other condiments**	1. **a person holding a bowl of oats next to bowls of other condiments** 2. <u>a meal is being displayed on a table</u> 3. **a table with bowls of grains and fruit and a hand with a plate** 4. <u>a table that has some food on it</u>
② 	1. a man riding a skateboard up the side of a ramp 2. a man riding a skateboard up the side of a ramp 3. a man at a skate park with his foot on the side of the skateboard 4. a man on a skateboard performing a trick	1. its a cloudy night for a ride on the motorcycle 2. **a motorcyclist surveys the sunlit road into the horizon** 3. **a close up of a person riding a motorcycle on a long empty road** 4. a photo taken from a car looking at a skateboarder on the side of the road	1. **a motorcyclist surveys the sunlit road into the horizon** 2. **a close up of a person riding a motorcycle on a long empty road** 3. a photo taken from a car looking at a skateboarder on the side of the road 4. its a cloudy night for a ride on the motorcycle
③ 	1. a couple of giraffes look around the ground in the zoo 2. two giraffe standing near brick building 3. a pair of giraffes standing around in their enclosure 4. two giraffes roaming around an enclosed area on a sunny day	1. a pair of giraffes standing around in their enclosure 2. **a couple of giraffes eating hay from a trough** 3. <u>two giraffes that are eating from a basket</u> 4. <u>two giraffes stand and eat food out of a basket</u>	1. **a couple of giraffes eating hay from a trough** 2. **a couple of giraffes eating out of a basket** 3. <u>two giraffes stand and eat food out of a basket</u> 4. <u>a couple of giraffes reach for a basket</u>

Note: groundtruth matched sentences are marked as **red**, while some sentences sharing similar meanings as groundtruths are marked as <u>underline</u>.

图 14：图像标注实例对比

2.2 记忆机制建模

王亮介绍的第二项工作是记忆机制建模，发表在 ACM MM 2019，该工作是关于视频描述 (Video Captioning) 任务的。

视频描述，即给定一个简短的视频，通过一句话描述视频当中发生的内容。它有很多潜在的应用，最典型是导盲：通过摄像头阅读道路环境，同时能够解说使用者周围的路况，可以协助盲人在公共场所的活动。

视频描述挑战有两个方面，一是如何在视觉空间和语言空间能够架起一个有效的映射关系，二是如何建模长时间的视觉与文本之间的依赖性。

众所周知，LSTM 在序列建模上有很好的效果，但是不能够很好地建模长期的依赖关系。而任务中的视频和相应的句子都是比较长的，比如大于 20 个点、30 个点以上的过程。王亮认为 LSTM 在这方面做得就不是很好，因此他们在工作中加入了记忆机制解决上述问题。

王亮团队提出的用于视觉描述的框架性网络如图 15 所示。图中最上面的模块是针对视觉的部分，使用 2D 或 3DCNN 来提取外部特征或运动特征； 最下面的部分是基于 LSTM 的文本解码器，用于生成文字。而架起视觉和文字之间的中间模块，就是所谓的记忆部分。在这个部分中，提出了三模块：TM 是指文本记忆，VM 指是视觉记忆，IM 指是属性记忆，分别存储基于 LSTM 文本解码器的表达，视频表达以及语义属性。

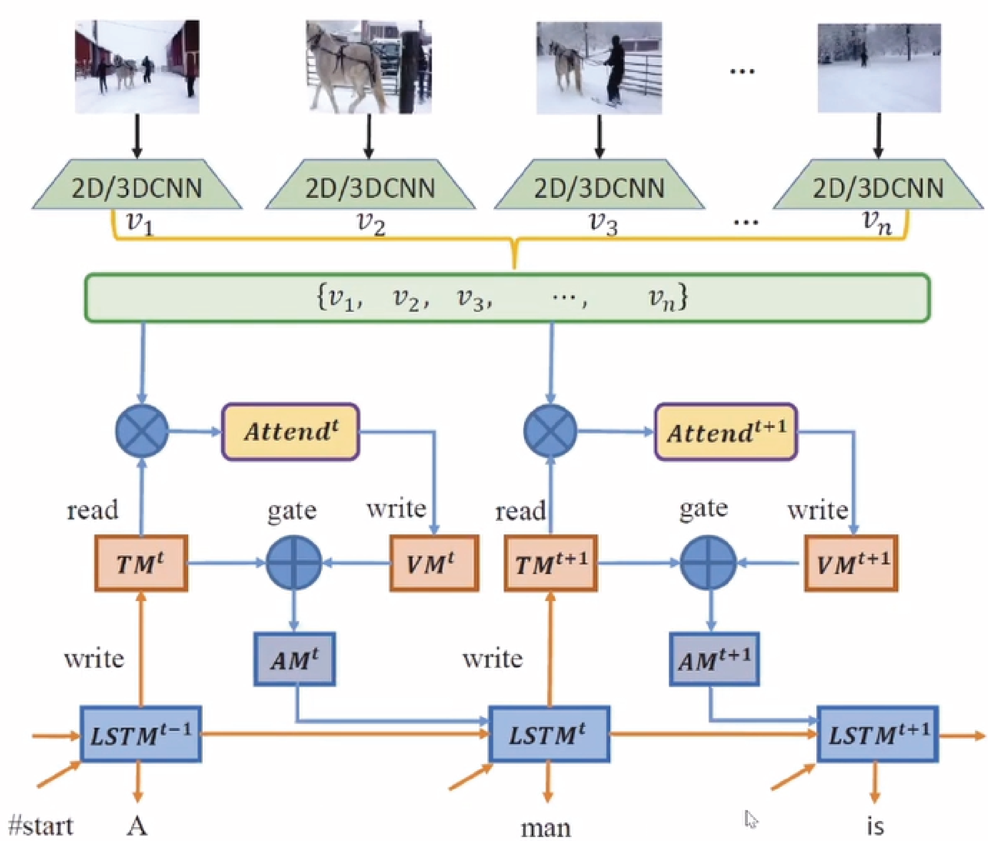


图 15：视觉描述的框架性网络

对于属性的检测来讲，王亮团队使用了下图所示的常见方法，

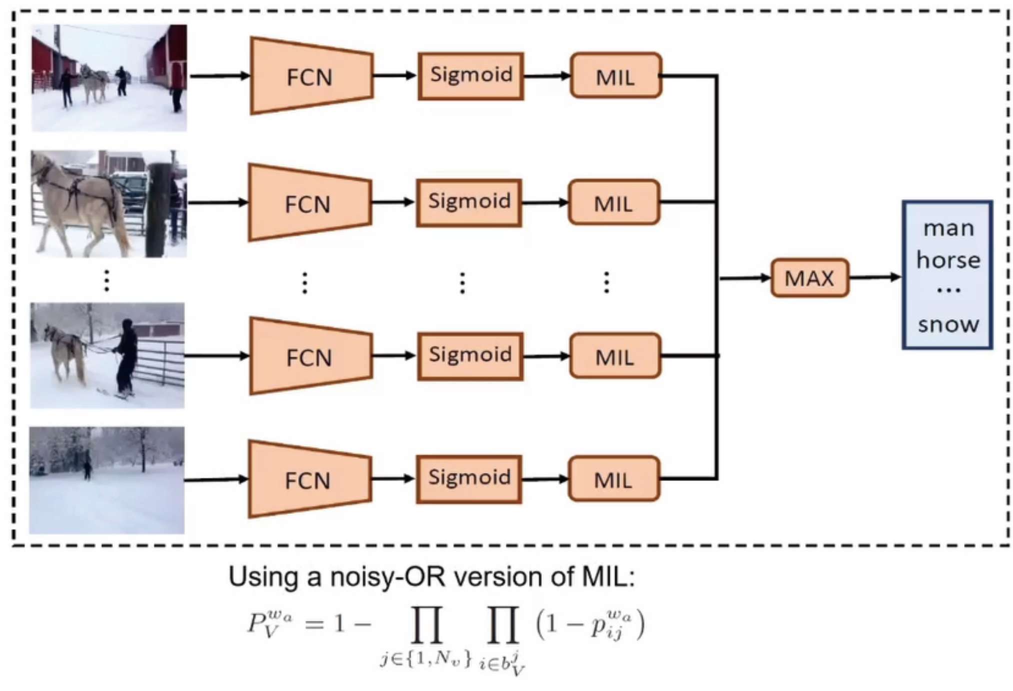


图 16：属性检测

如上所述，记忆建模采取分层的记忆建模机制，提取了三个方面的记忆：文本的记忆；属性的记忆；视觉的记忆。三个方面之间进行如下的交互，完成视觉描述的任务。

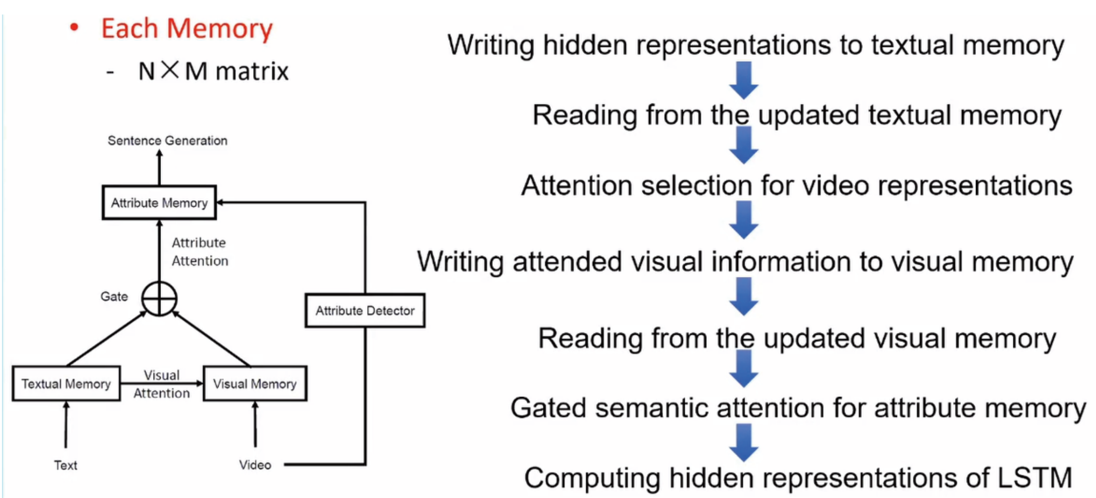


图 17：分层的记忆建模机制

将上述模型应用在几个主要数据集上，结果如图 18 所示。王亮团队所提的分层视觉记忆机制是在几个数据库上面获得了最好的性能，并且该模型框架对于不同视觉特征的输入是不敏感的。

Video captioning results on MSVD dataset.

Method	BLEU@4	METEOR	CIDEr-D
FGM [25]	13.68%	23.90%	-
LSTM [30]	33.29%	29.07%	-
SA [39]	41.92%	29.60%	51.67 %
S2VT [29]	-	29.8%	-
LSTM-E [15]	45.3%	31.0%	-
p-RNN [42]	49.9%	32.6%	65.8 %
HRNE [14]	43.8%	33.1%	-
BGRCN [1]	48.42%	31.70%	65.38
HBA [2]	42.5%	32.4%	63.5
RMA [12]	45.7%	31.9%	57.3
TSA [16]	52.8%	33.5%	74.0
HMM-c3d	46.0%	31.8%	63.2%
HMM-vgg	48.7%	32.4%	67.6%
HMM-inv3	49.9%	33.0%	70.7%
HMM-c3d-vgg	50.8%	33.1%	72.2%
HMM-c3d-inv3	**52.9%**	**33.8%**	**74.5%**

Video captioning results on MSR-VTT dataset.

Method	BLEU@4	METEOR	CIDEr-D
EMLR [8]	38.7%	26.9%	**45.9**
MVD [19]	38.3%	27.0%	41.8
Aalto [21]	39.8%	26.9%	45.7
MP-vgg [30]	34.8%	24.8%	-
SA-vgg [39]	35.6%	25.4%	-
SA-c3d [39]	36.1%	25.7%	-
SA-c3d-vgg [39]	36.6%	25.9%	-
SA-inv3 [39]	36.3%	23.6%	37.7%
SA-c3d-inv3 [39]	38.1%	25.9%	38.0%
HMM-c3d	37.4%	26.2%	40.5%
HMM-vgg	36.8%	26.0%	38.6%
HMM-inv3	37.2%	26.5%	39.4%
HMM-c3d-inv3	**39.9%**	**28.3%**	40.9%

图 18：实验结果

此外，王亮给出了一些定性化的例子，SA 作为比较的方法，HMM 是所提方法。以表格中第二行为例，给定一个输入的视频，SA 的输出结果是“A man is playing with a guitar”，而 HMM 的结果是“A man is playing with a dog”，而视频中本身就是一个狗，不是吉他。可见在生成句子的语义之前，所提 HMM 方法可以得到更准确的视觉字。

SA：Yao et al. ICCV 2015　　HMM：Our model

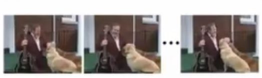
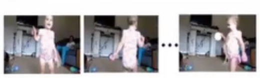
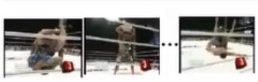
	Attributes from Video: Ball, teams, players, game, people, men, player, soccer, football, playing	Generated Sentence: SA: people are playing HMM: a group of men are playing soccer	Reference Sentece: 1. two teams are playing soccer 2. men are playing football 3. some men are playing soccer
	Attributes from Video: playing, going, person, moving, man, dog, front, funny, white, cute	Generated Sentence: SA: a man is playing a guitar HMM: a man is playing with a dog	Reference Sentece: 1. a man is petting two dogs 2. a man pets some dogs 3. a man is play with pets
	Attributes from video: child, little, baby, kid, toddler, cute, playing, using, trying, high	Generated Sentence: SA: a baby is playing on a couch HMM: a baby is playing with a toy	Reference Sentece: 1. a baby is playing with toys 2. a toddler is picking up toys 3. the baby is putting away her toys
	Attributes from video: pan, woman, cooking, pot, preparing, frying, person, something, mixture, bowl	Generated Sentence: SA: a woman is cooking the kitchen HMM: a woman is cooking some sort of food in a pan	Reference Sentece: 1. a lady is cooking food 2. the lady prepare the food 3. a person making nabeyaki udon noodle
	Attributes from video: Man, his, playing, animal, boxing, lady, black, person, around, doing	Generated Sentence: SA: the monkey is playing a monkey HMM: two men are boxing in a boxing ring	Reference Sentece: 1. two men are boxing in a ring 2. two men fight inside a ring 3. two persons are fighting

图 19：实例对比

2.3 推理机制建模

王亮介绍的最后一个工作是推理机制的建模，发表于 CVPR 2019 oral，与行为定位相关。

王亮首先解释了行为分类、行为定位的基本概念。行为分类指给定一段裁减好的视频，标注该视频的类别。而行为定位在行为分类的基础上更进一步，不光给行为做标注，同时也要标注行为在一段长视频中间的起始位置。具体实例如下图所示。行为的定位对视频监控十分意义，因为视频监控通常都是非常长且没有裁减，如果想在视频监控中找到一个特定的事件起始时间点，行为检测就是非常重要的环节。

在这项工作中，王亮团队设计了面向语言驱动的视觉行为定位，架起了视觉与语言的桥梁。

语言驱动的视觉行为定位，即给定一个语言的查询，查找视频中对应行为发生的起始时间点。图 20 中，“一个人走下了窗口，并且向外看”，就是一个语言的查询事件。这对视频监控是非常重要的过程，因为在实际事件中，行为通常是非常复杂的，多变的，不可能通过一个单一的字进行描述，所以通过语言的查询去检索事件是非常重要。

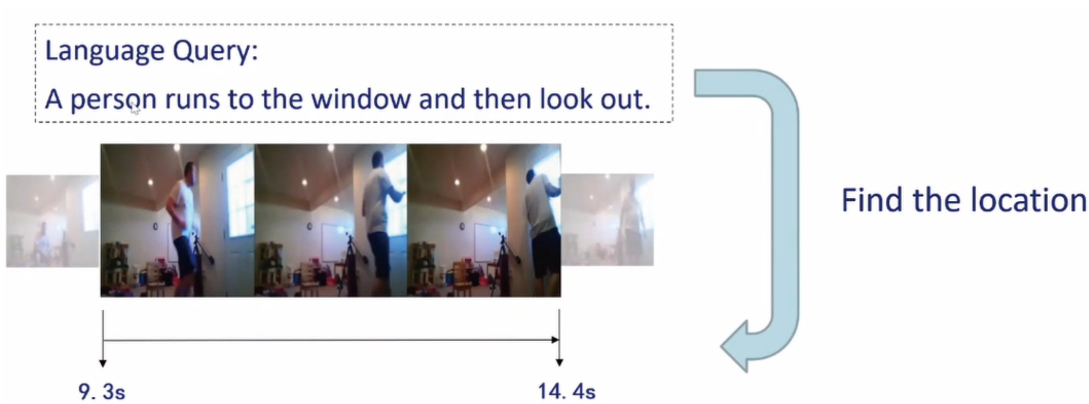


图 20：查找视频中对应行为发生的起始时间点

传统的解决方法，如图 21 所示，也是对语言和视频提取全局特征，然后通过不同的损失的来解决行为检测的问题。但王亮团队认为，当前的很多方法在时间上计算复杂度非常高，因为采用滑动窗的形式在长视频中不断地去试，直到找到目标的行为，这是非常耗时的。并且在此过程中时间的信息没有被完全充分地探索。

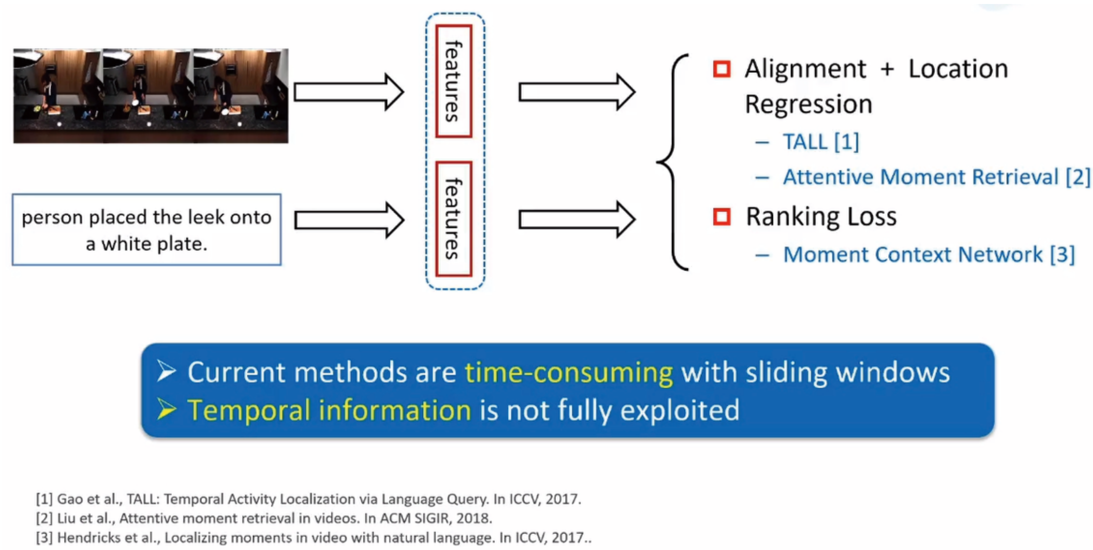


图 21：相关工作

王亮团队提出了语义匹配的强化学习网络，通过这样的网络选择性地观察一序列视频，从而发现视觉语义的相关信息。相关信息即为对应所查询的语义事件。

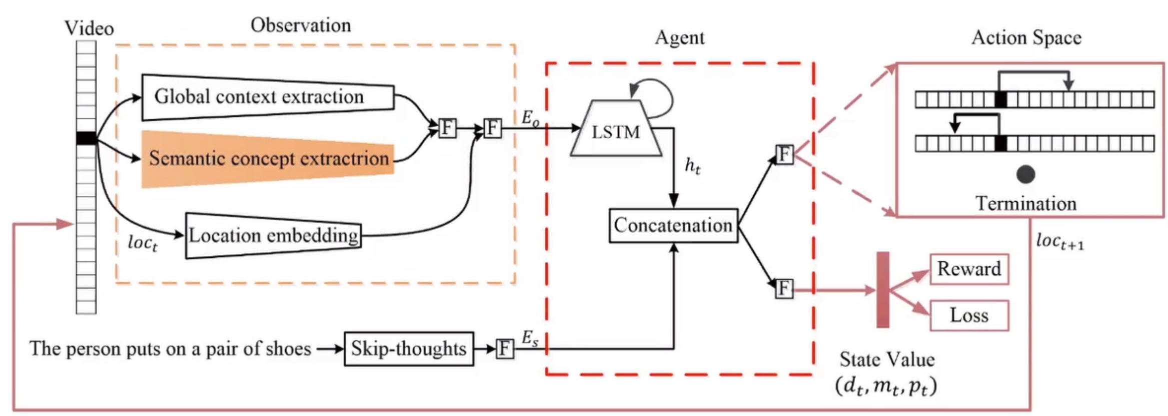


图 22：语义匹配的强化学习网络

具体来说，该模型所用强化学习的代理基于 LSTM，因为 LSTM 可以动态地去观察一序列的视频任务，并且最终输出所检测的结果。在前向通路中，句子的查询是通过 skip-thoughts 编码，同时全局的上下文特征和语义的概念特征也被提取出来，所用方法与前文介绍的内容相同；当前观察帧的位置也被嵌入特征中。接下来是 LSTM 序列的总结信息，也是从历史的观察进行信息总结，并且编码视频的时间信息。隐藏的状态和句子结合在一起，最终输出行为和状态值，行为用来选择下一个观察的位置，状态值（State Value）包含一个候选的检测结果、一个匹配的分值，还有一个二值化预测的指示器。奖励或惩罚的函数通过状态值加以计算。

$$L(D) = \omega_1 \sum_t L_{cls}(m_t) + \omega_2 \sum_t L_{loc}(d_t, g_t) \begin{cases} L_{cls}(m_t;\theta_m) = -\sum_i (r_i) log P(r_i|m_i;\theta_m) \\ L_{loc} = \frac{1}{n}\sum_i^n \sum_j^M [x_{ij}\log(p_{ij}) + (1-x_{ij})\log(1-p_{ij})] \end{cases}$$

$$r_T = \begin{cases} R_{FN}, & FN \\ N_{TP}R_{TP} + N_{FP}R_{FP}, & TP \text{ and } FP \end{cases}$$

上述公式给出了具体的损失函数的形式和语义概念提取的形式。

王亮团队在当前的典型数据集上做了行为定位的实验，结果如图 23 所示。从实验结果可以看出，所提方法的速度是传统方法速度的 6 倍以上，并且语义概念的嵌入确实带来了性能的提高。

Table 1: Results on TACoS and Charades-STA datasets

Method	TACoS R@1 IoU=0.5	TACoS R@1 IoU=0.3	TACoS R@1 IoU=0.1	TACoS R@5 IoU=0.5	TACoS R@5 IoU=0.3	TACoS R@5 IoU=0.1	TACoS mR	Charades-STA R@1 IoU=0.5	Charades-STA R@1 IoU=0.7	Charades-STA R@5 IoU=0.5	Charades-STA R@5 IoU=0.7	Charades-STA mR
Random	0.83	1.81	3.28	3.57	7.03	15.09	5.27	8.51	3.03	37.12	14.06	15.68
CTRL [8]	13.30	18.32	24.32	25.42	36.69	48.73	27.80	23.63	8.89	58.92	29.52	30.24
RL(b)	11.76	17.70	22.42	22.61	33.24	45.10	25.47	19.78	5.60	55.65	25.07	26.53
RL(f)	12.79	18.53	23.87	24.56	35.30	47.64	27.15	21.18	7.33	56.01	27.85	28.09
SM-RL(attr+b)	13.50	18.83	23.72	24.01	34.19	46.56	26.80	21.00	7.63	57.25	28.06	28.49
SM-RL(attr+f)	14.01	19.02	23.96	24.55	36.42	47.14	27.51	22.54	8.56	58.95	29.74	29.95
SM-RL(attr*+b)	14.20	19.79	25.17	25.38	36.69	48.22	28.24	23.56	9.52	60.17	32.53	31.45
SM-RL(attr*+f)	**15.95**	**20.25**	**26.51**	**27.84**	**38.47**	**50.01**	**29.84**	**24.36**	**11.17**	**61.25**	**32.08**	**32.22**

Table 2: Results on DiDeMo dataset

Method	Rank@1	Rank@5	mIoU
MCN([11])	28.10	78.21	41.08
SM-RL(attr*+b)	29.64	79.38	42.17
SM-RL(attr*+f)	**31.06**	**80.45**	**43.94**

Table 3: Detection speed comparison

Method	Average running time (per minute video)
CTRL	202ms
Ours	**32ms**

图 23：行为定位实验结果

进一步地，王亮给除了两个定性的例子。如图 24 所示，例 1 查询的是“the person washes the leeks in the sink”，即一个人在水池当中淘洗韭菜，可以看到，当代理接近于达到行为结束时间时，它往后倒退了一步，来优化假设的位置，例 2 查询的是“Person put on a pair of shoes”，即一个人穿上了一双鞋，从中也可以看处，对于这个句子的查询，整个系统在结束前向后倒退了两步，以优化起始时间和结束时间。

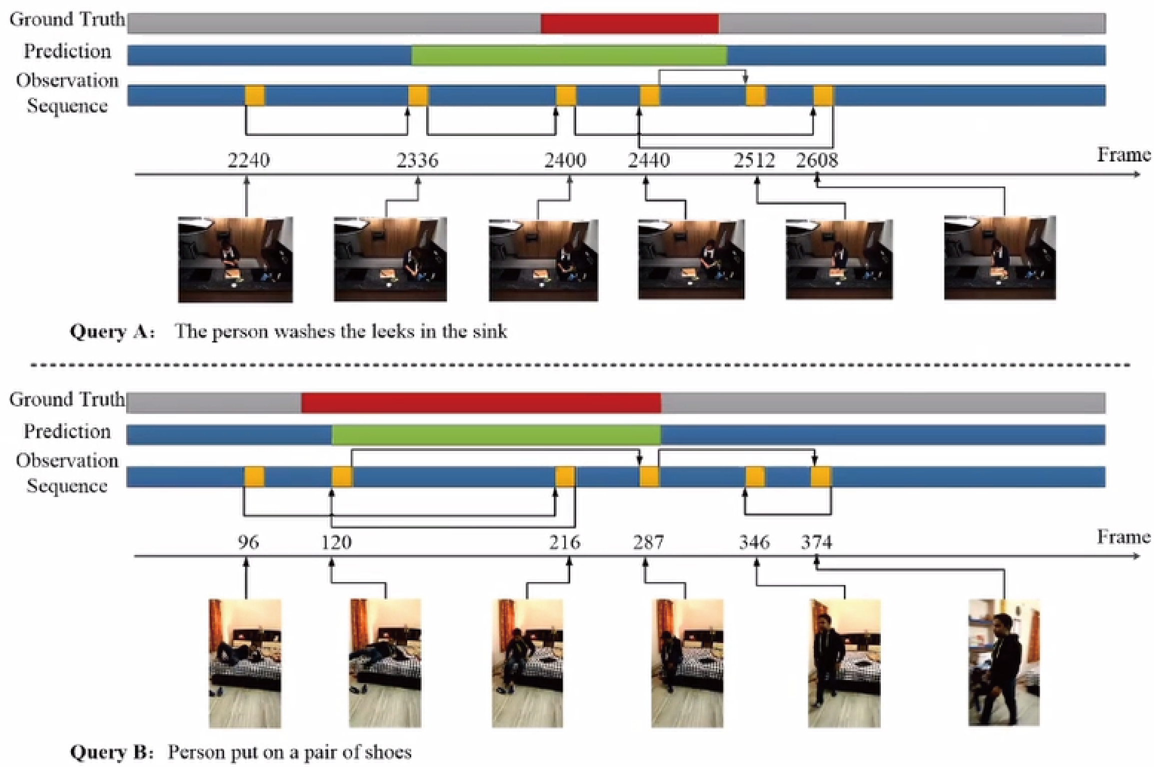


图 24：定性的例子

三、总结与未来工作

最后王亮对上述工作做一个简单的总结：

- 注意建模可以选择性地处理视觉信息，减少冗余信息的影响；
- 记忆建模能够存储历史信息或者先验知识，且能够再使用表达没有看到的或小样本的信息；
- 推理建模能够捕捉到目标属性等之间的潜在关系，支持在高层做出判断。
- 整个复杂视觉任务的处理过程中，学习中级的视觉属性能够帮助减少大的视觉语义的鸿沟。

同时王亮也从三个方面简单地介绍了未来的工作方向：

- 如何做 Decision Making。虽然注意、记忆、推理的建模有助于提升复杂视觉任务的性能，但做决策是最终目的。目前的工作与决策相关的不多，有也大多是都是单个步骤的决策，研究进行更深一步的多步决策是比较有意义的事情。
- 视觉、语言和音频的结合。当前更多的模态的识别都是单一处理，比如视觉、语音或者音频；跨模态的检索和分析都是两两模态之间。但其实每一种模态之间都有互补性的作用，在视觉语言的基础上再增加一个音频也是需要研究的问题。
- 与类人智能机器人结合。类人智能机器人是视觉研究最容易应用的载体，把复杂的任务或者视觉与外界的交互这样的分析和与类人智能机器人结合有实际应用价值。

下图为近几年来王亮团队在多模态学习相关发表的文章。

- Niu et al., Improving description-based person re-identification by multi-granularity image-text alignments, *IEEE TIP*, 2020.
- Wang et al., Long video question answering: A Matching-guided Attention Model, *PR*, 2020.
- Wang et al., Learning Visual Relationship and Context-Aware Attention for Image Captioning, *PR*, 2020.
- Huang et al., Image and Sentence Matching via Semantic Concepts and Order Learning, *IEEE TPAMI*, 2019.
- Yin et al., Multi-view Clustering via Unified and View Specific Embeddings Learning, *IEEE TNNLS*, 2018.
- Wang et al., Joint Feature Selection and Subspace Learning for Cross-modal Retrieval, *IEEE TPAMI*, 2016.
- Jia et al., Cross-Modal Cross-Domain Moment Alignment Network for Person Search, *CVRP*, 2020.
- Huang et al., ACMM: Aligned Cross-Modal Memory for Few-Shot Image and Sentence Matching, *ICCV*, 2019.
- Huang et al., Few-Shot Image and Sentence Matching via Gated Visual-Semantic Embedding, *AAAI*, 2019.
- Huang et al., Learning Semantic Concepts and Order for Image and Sentence Matching, *CVPR*, 2018.
- Wang et al., Language-driven Temporal Activity Localization: A Semantic Matching RL Model, *CVPR*, 2019.
- Wang et al., Stacked Memory Network for Video Summarization, *ACM Multimedia*, 2019.
- Wang et al., M3: Multimodal Memory Modelling for Video Captioning, *CVPR*, 2018.
- Li et al., Cross-Modal Ranking with Soft Consistency and Noisy Labels for Robust RGB-T Tracking, *ECCV*, 2018.
- Wang et al., Hierarchical Memory Network for Video Captioning, *ACM Multimedia*, 2018.
- Huang et al., Instance-aware Image and Sentence Matching with Selective Multimodal LSTM, *CVPR*, 2017.

图 25：相关文章

北大教授吴玺宏：一种具身自监督学习框架——面向任何语种的音系构建任务

整理：智源社区 蒋宝尚

在第二届北京智源大会“机器感知论坛”上，北京大学信息科学技术学院副院长吴玺宏教授做了《一种具身自监督学习框架——面向任何语种的音系构建任务》的报告。

在报告中，吴玺宏提到，所谓“具身自监督学习框架”指的是就是利用人的身体和物理的模型帮助自监督学习。这种具身自监督学习框架能够克服传统方法存在的很多问题，例如数据获取代价太大，同样存在推广性的问题等等。

这种方法旨在对任意方言、语种的发音姿态和音系构建，从而辅助语音学家以及语言教学工作，也可以推动方言、小语种的语音识别和语音合成的研究。此外，基于物理模型或实际物理过程的具身学习方法，可推广到视觉 – 运动、听觉 – 运动等机器感知工作。

以下是智源社区编辑对吴玺宏演讲进行的文字整理:

我的报告主题是“一种具身自监督学习框架，面向任何语种的音系构建任务”，顾名思义，这种具身自监督学习框架，目的在于思考当面向任意语种和方言时，能否构建出音系？例如拼音系统。如果能够构建出音系，那么就能实现任何方言、语种的识别。

所谓“具身自监督学习框架”指的是就是利用人的身体和物理的模型帮助自监督学习。在介绍这个模型之前，先看语音识别的研究现状、困境和反思。

一、语音识别的研究现状、困境与反思

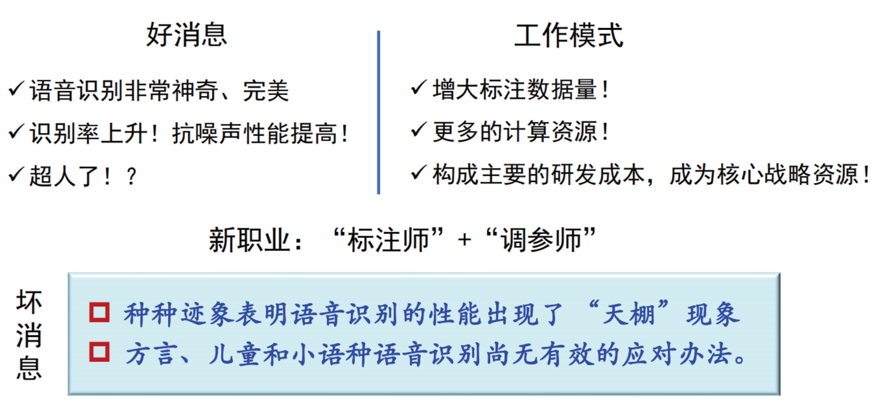


图 1： 语音识别的现状和困境

当前的语音识别研究现状好的一面，最直观的展现是：识别率上升，抗噪声性能提高。但是识别率上升，需要付出的成本是增大标注数据量、更多的计算资源。因此，数据和计算资源慢慢构成企业主要的研发成本，成为核心战略资源。

而当前语音识别的缺陷和挑战有两个方面：其一，随着标注数据量增大，种种迹象表明语音识别的性能出现了“天棚”现象；其二，方言、儿童和小语种语音识别尚无有效的应对办法。

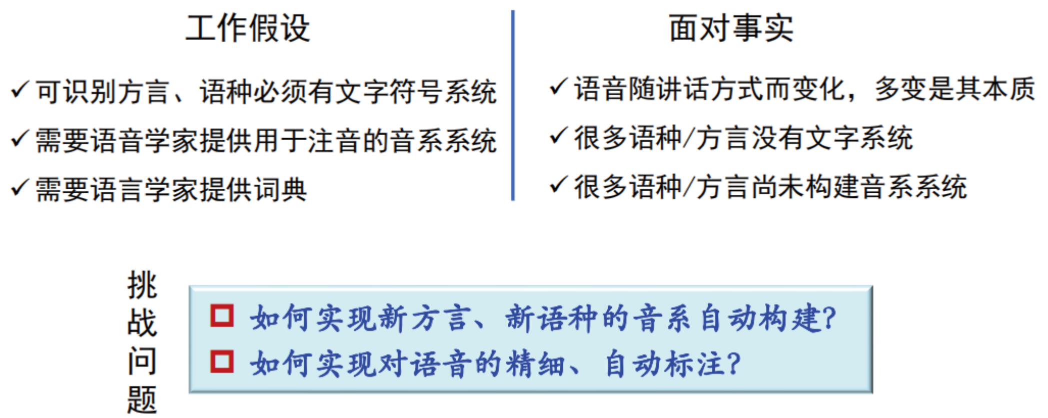


图 2：反思：语言识别研究隐含的工作假设

为什么会出现上述两个挑战呢？经过反思，发现其原因可能是，语音识别研究隐含的工作假设和现实的不一致造成的。

当前语音识别研究隐含的工作假设主要有三个，首先，可识别方言、语种必须有文字符号系统，例如东北方言中的“波棱盖”一词，必须在字典中有明确的文字符号（虽然有的方言有文字系统，但是不全面）；其次，除了文字符号之外，还需要语音学家提供用于注音的音系系统；最后，需要语言学家提供词典。只有满足上述三个假设，语音识别工作才能完成。

编者注：波棱盖音为“bō lēng gài er”，书面意思是膝盖。造句为：波棱盖儿卡马路牙子上秃噜皮了。秃鲁皮指皮肤表面被擦破或被开水烫过的状态，马路牙子是马路边上，这句是膝盖碰在马路边上把皮擦破了。

而现实是：1. 语音可能随着讲话的方式而变化，多变已经成为语音的本质；2. 许多语种 / 方言其实并没有文字系统；3. 很多语种 / 方言尚未构建音系系统。

因此，“假设”与“事实”不一致，便带来了两个挑战问题，即如何实现新方言、新语种的音系自动构建？如何实现对语音的精细、自动标注？下面我介绍具身认知启发的表示学习与范畴学习，从而尝试找到解决这两个问题的方案。

二、具身认知启发的表示学习与范畴学习

首先回顾一下语音产生的过程：嘴通过肺部的气流压力，经过声门激励到咽腔、口腔和鼻腔。这期间经历的姿态变化会产语音。随后，语音会经过人耳，在语谱的作用下完成听觉过程。

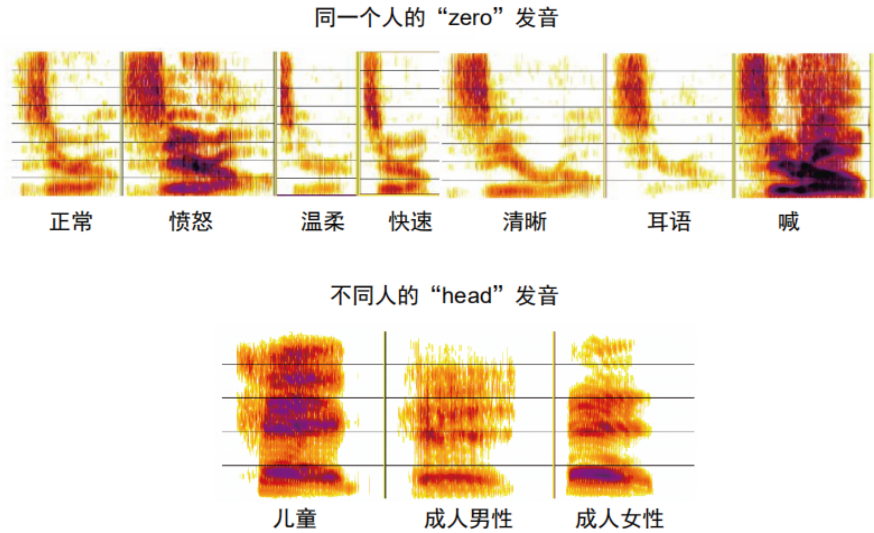


图 3：语音的多变性

此外，语音具有有多变性，例如同一个人对单词“zero”进行发音，会得到不一样的效果，例如：同一个人的正常发音、温柔发音、快速发音都会产生不一样的效果；吐字清晰、小声耳语以及喊叫的语谱的变化非常之大。另外，不同人的“head”发音，对应的成年男性、成年女性的语谱差别也很大。

所以，基于上述观察，如果想通过语谱的表达把声音信号进行范畴化和分类，便会出现推广性问题，即语音语谱表示的多变性，不可能覆盖所有的变化数据！这个推广性问题产生“数据越多好”的现象，但同时也要明确，在语音层面，有永远都见不到的数据。

同样，也会出现可解释性问题：人工粗略标记，不能描述精细变化！例如只能对着一段语谱标注一个符号，没有进行共振风和谐波的标注，无解释“音”如何发出来。因此，通过语谱表达语音似乎行不通。

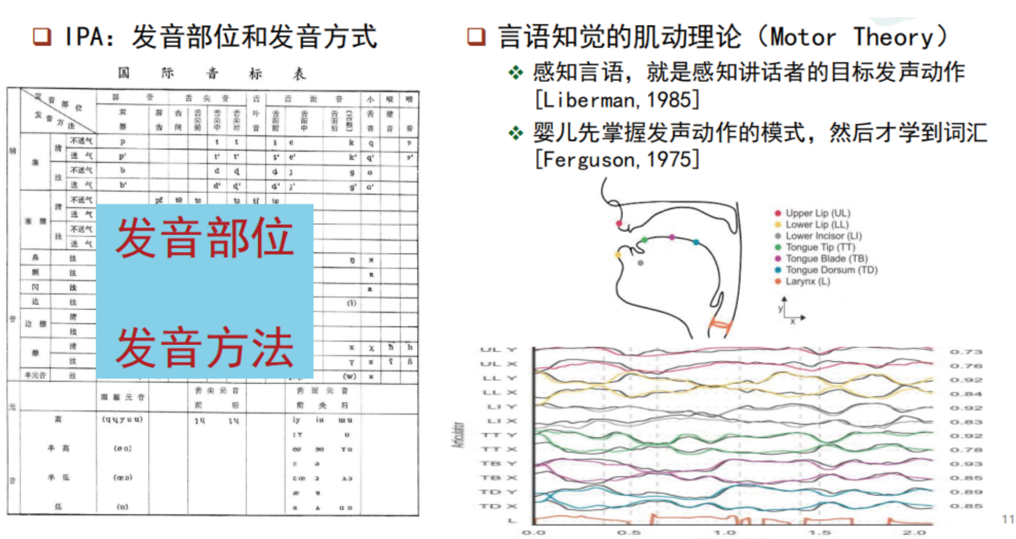


图 4：基于肌动理论的音位系统构建

在这种情况下，我们便探索了基于肌动理论的音位系统构建，采用了国际音标表和言语知觉的肌动理论两个“辅助工具“，其中，国际音标（缩写：IPA），是一套用来标音的系统，以拉丁字母为基础，由国际语音学学会设计来作为口语声音的标准化标示方法，能够把世界各地的语言语音音位刻画出来。

言语知觉的肌动理论有两个重要的结论，一个是在 1985 年，Liberman 提出的，即感知言语，就是感知讲话者的目标发声动作； 另一个是 1975 年，Ferguson 提出的，即婴儿先掌握发生动作的模式，然后才学到词汇。显然，这两个理论都表达了发声姿势的重要性。

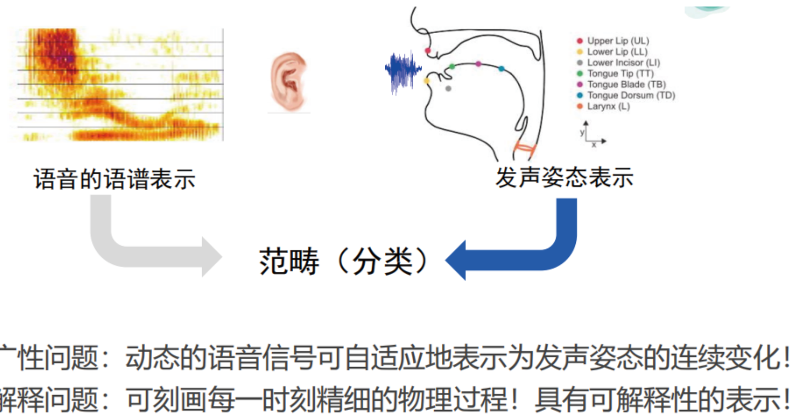


图 5：表示问题和范畴问题

因此，我们如果听到语音不是经过语谱表达，而是通过语音推测发言姿态，是不是就可以解决推广性的问题？其实，通过语音推测发言姿态不仅能够解决推广性问题，还能够解决可解释问题。因为，动态的语音信号可自适应地表示为发声姿态的连续变化，意味着可以刻画每一时刻精细的物理过程。

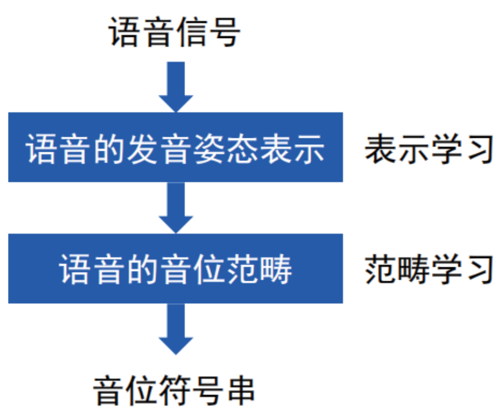


图 6：表示学习和范畴学习

针对表示问题和范畴问题，进行的表示学习和范畴学习的过程如上图所示。首先通过语音信号，从而进行语音的发音姿态表示（表示学习），然后“进入”语音的音位范畴（范畴学习），就能够得到音位符号串（对某种方言构建音系系统）。

三、基于发声物理模型的语音发声姿态：自监督学习

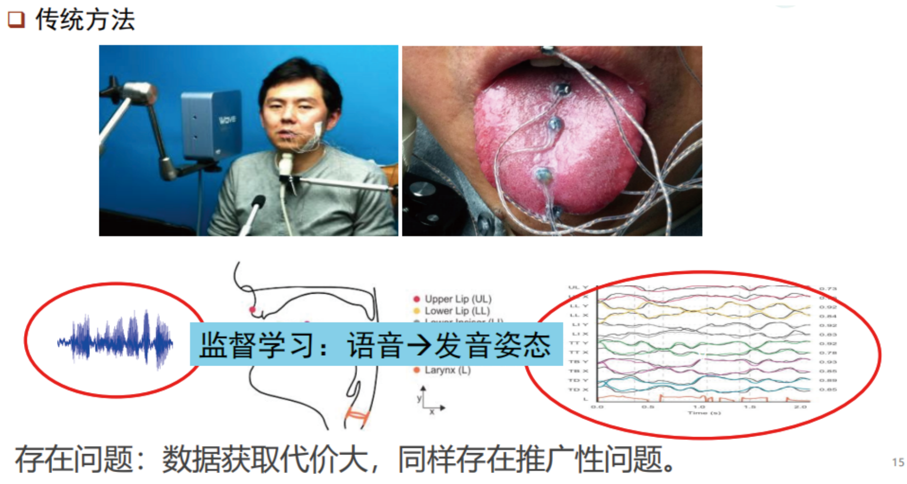


图 7：从语音到发音姿态

关于发音姿态预测其实已经有了不少研究，传统的方法叫语音反演，即从语音到发音姿态。具体操作过程如上图所示：通过在舌部贴上传感器，经过外部的设备可以捕捉发声过程当中口腔的姿态，如此便能实现发音的时候记录语音和各个姿态，即得到语音数据和发音姿态的数据，然后经过监督学习实现语音到发音姿态的反演。

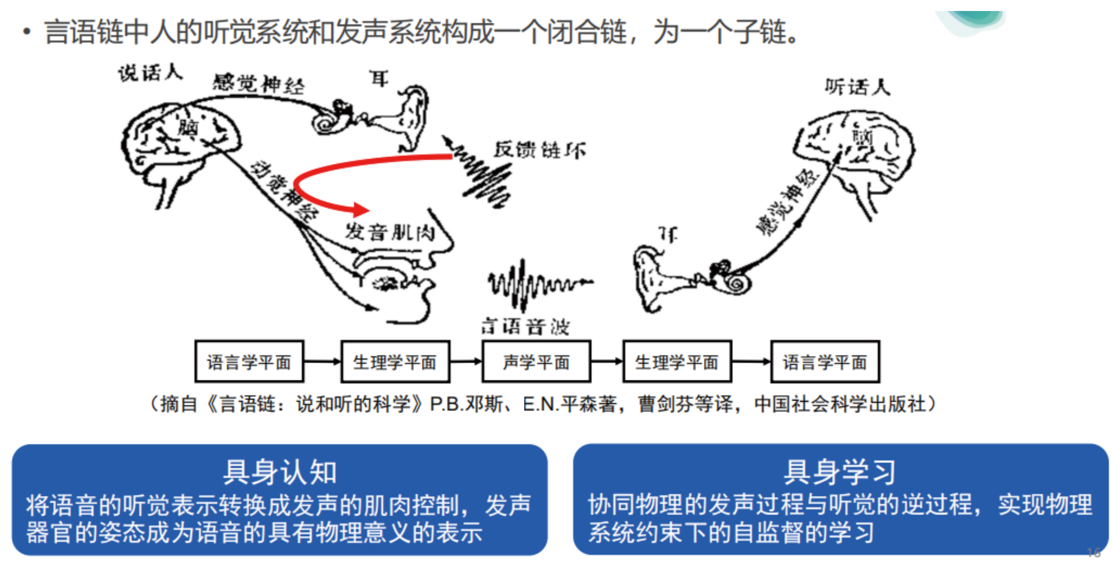


图 8：具身认知与具身学习

传统方法存在很多问题，例如数据获取代价太大，同样存在推广性的问题。如何解决？首先要明确听觉系统是存在一个言语链的概念，即通过神经驱动肌肉进行控制说出语音，然后进入耳朵解码，解码之后，还需要耳朵进行校正校准。

因此，自己发声、自己听，就会反馈构成一个闭合链。根据闭合链，我们提出具身认知概念，即将语音的听觉表示转换成发声的肌肉控制，发声器官的姿态成为语音的具有物理意义的表示； 具身学习概念，即协同物理的发声过程与听觉的逆过程，实现物理系统约束下的自监督的学习。

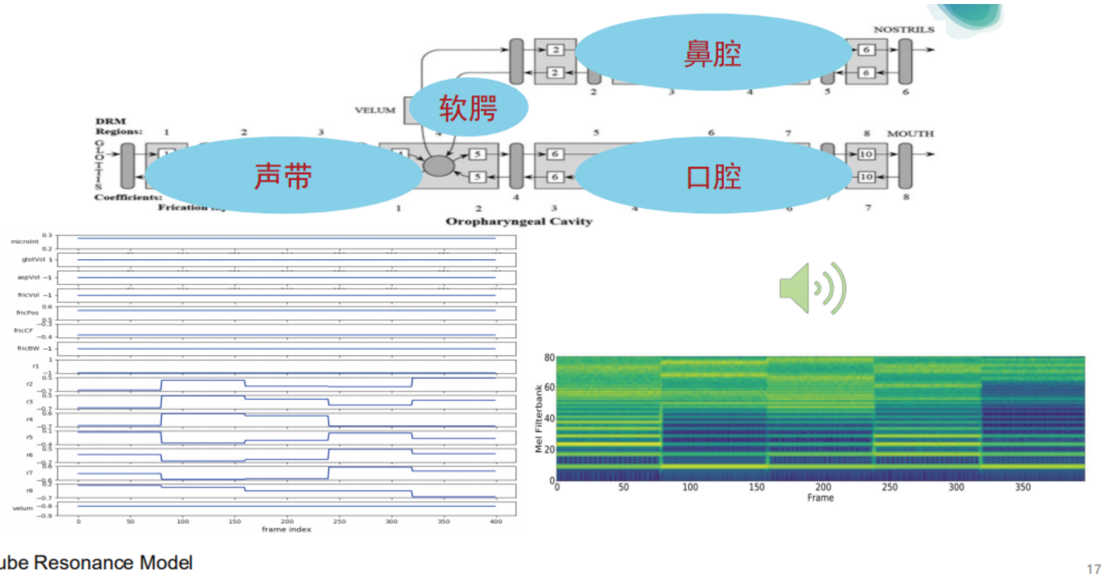


图 9：声门 – 声道发声滤波器 TRM 模型 [Hill,2017]

有了概念，具体怎么操作呢？在 2017 年，有一个声门 – 声道发声滤波器 TRM 模型，包含的要素包括： 声门、声道、鼻腔、口腔以及（小舌）软腭的调节。如上图所示，此物理模型能够通过这些短的声管进行串联和并联，从而控制这个气流激励、声带振动，得出声音。

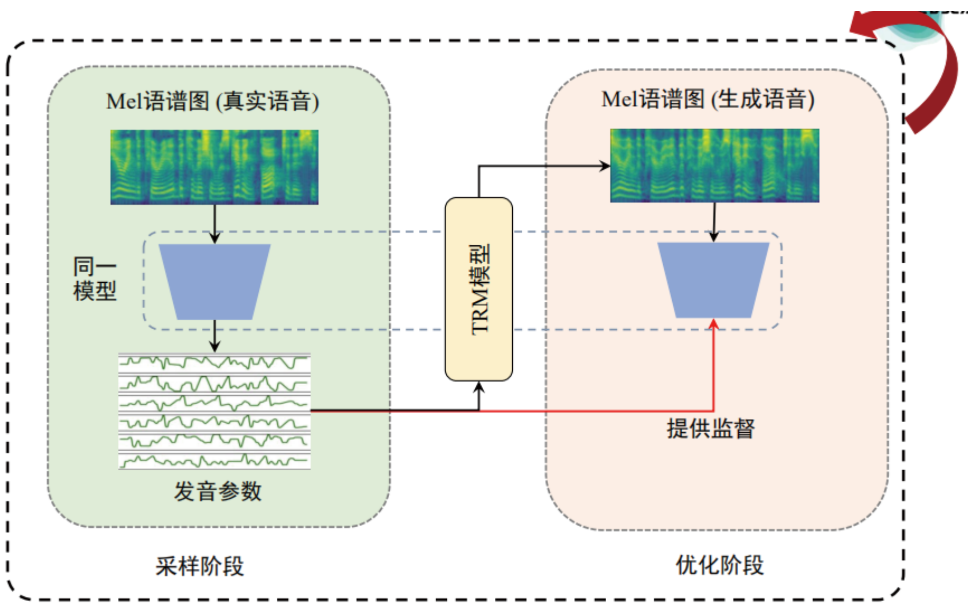


图 10：具身自监督学习框架

有了物理模型之后，我们又提出了具身自监督的学习框架：首先搜集语音信号，然后通过神经网络推测发音参数，随后用发音参数驱动 TRM 模型。由于姿态是不准确的，发出的声音也是不准的，但是此声音一定是物理模型产生，于是会和“不准确的姿态”有对应关系。随后把不准确的声音信号经过神经网络进行姿态推测学习，即作为监督信息求解梯度方向，实现一轮的学习。

如此循环反复，经过不断的调整和迭代，就能够实现具身基于 TRM 物理模型的自监督学习框架。

效果如何呢？我们采用 10 个小时，单说话人（英语、女性）的语音库，经过网络模型的迭代学习，能够推特测出发音姿态。另外，也可以合成语音。

在汉语层面上，模型的推测结果是：汉语的辅音、元音比较准确，声调、韵律和声调都能恢复出来。

因此，通过物理模型的帮助，能够实现语音到发音姿态的表示和转换。在此基础之上，我们希望能够建立一个音位范畴，如此便能对任意语音（方言）进行识别，

四、音位范畴学习及音系构建

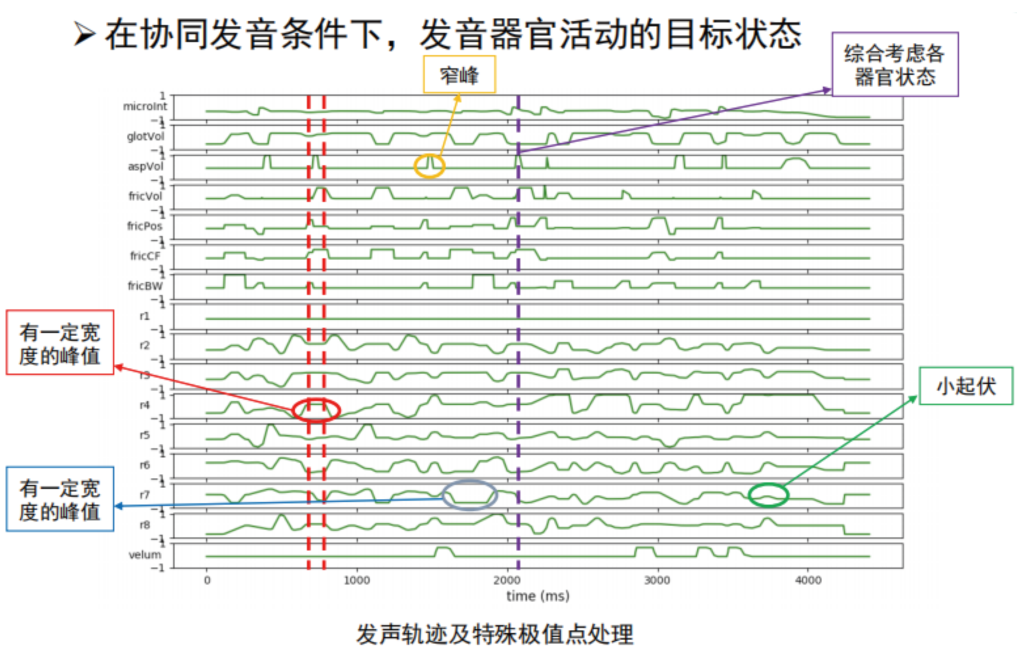


图 11：发音目标姿态

音位范畴怎么构建？虽然发音的姿态随着时间一直在动态变化，但是发音过程中，每个音位都有一个目标，这些“目标“会有一些特点？例如一定的峰值和宽度，当然也有一些窄峰的情况。

所以在实际构建的过程中，需要考虑各个器官和部位之间协同的变化，以及实际的语音信号，自适应汉语的信号，从而在语谱当中推测出来发音姿态，从中继续推测这些发音的目标。

找到“目标”之后，然后可以进行降维处理，包括采用上下文加权等方法，把这些所谓的“目标”进行聚类，从而形成范畴分类。

形成范畴之后，便可以进行语音识别，毕竟这相当于建立一套新的拼音系统。此外，有了语音就可以实现文字符号的转换。

五、结论和展望

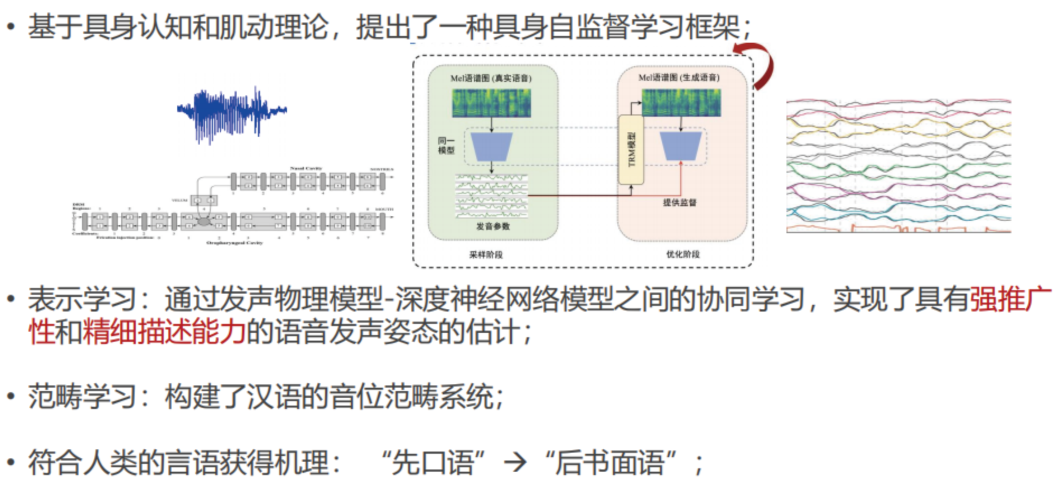


图 12：一种具身自监督学习框架

最后做出一些结论和展望：首先我们的工作基于具身认知和肌动理论，提出了一种具身自监督学习框架；然后我们进行了表示学习的工作，通过发声物理模型 – 深度神经网络模型之间的协同学习，实现了具有强推广性和精细描述能力的语音发声姿态的估计；此外，还通过范畴学习，构建了汉语的音位范畴系统；最后，通过实验证明可以构建符合人类言语获得机理的系统。

展望一下：我们的工作希望能够对任意方言、语种的发音姿态和音系构建，其作用可以辅助语音学家以及语言教学工作，也可以推动方言、小语种的语音识别和语音合成的研究。

此外，基于物理模型或实际物理过程的具身学习方法，可推广到视觉 – 运动、听觉 – 运动等机器感知工作。通过重新定义机器学习，期待解决推广性、泛化问题。也希望能够通过机器智能，将我们的身体与智能系统紧密联系在一起，将身体做为智能的来源。

最后，需要重新考虑未来智能研究的核心战略资源是什么？是标注大数据？还是其它资源？这个问题供大家思考。

问答

观众： 如果能够从语音获取发言姿态，是否可以考虑由于不同的人具有不同的发言习惯，不同人同一句话发言姿态实现语音识别或者活体检测？

吴玺宏： 我们的第一步是表式学习，不同的人说同样的话，发言姿态一定是不一样的，现在能够忠实地从语音信号当中找出发言姿态，然后就可以进行后续的说话识别以及活体检测。就像我们说的声源一样，光源打到脸上再反射回来就是一个物理过程，再就是人脸肌肉拉伸使得面部表情发生变化，这也是一个物理过程。实际上很多工作都有物理过程，物理模型以后可能是结合我们所谓的学习开辟很多新的天地。

观众： 语音方面有没有可能通过面部图象处理或者面部运动机理让聋哑人发声?

吴玺宏： 关键要有一个信息源驱动，如果只是因为声带问题，比如喉部做手术了，嘴如果能动，就能通过这个模型完全是可以出声音的，如果嘴不知道怎么动，那我也不知道怎么发音了。

观众： 声音的物理模型是不是很重要?

吴玺宏： 没错，物理模型太重要了，就是因为有物理模型才不用监督标签。

清华大学教授孙富春：机器人视触觉传感器与主动感知

整理：智源社区 张鲁洋

在第二届北京智源大会“机器感知”专题论坛上，清华大学孙富春教授做了《机器人视触觉传感器与主动感知》的报告。

机器人如何能够做到心灵手巧呢？一个非常重要的部分就是感知，包括触觉视觉实现对目标的精确感知和理解。而且还要像人一样能够学习技能。

我们国内的触觉传感器达到了什么水平？国内的触觉传感器目前也有“**卡脖子**”的问题，关键零部件目前还依赖国外进口，国产传感器在稳定性和一致性方面有待加强，如量程问题、材料问题，但是国内目前也在加强这方面的工作。

如何提高机器人的感知能力？一是“**传感器要长腿，即主动感知**”； 二是要会**选择**合适的传感器； 三是让机器人的传感器不断地**增长知识和发育**。

通过孙富春的报告，我们会了解到他团队已经开发了很多性能较好的传感器，比如压阻型柔性触觉阵列传感器、电容式传感器、指尖传感器等，一些在国际上达到先进水平。这些传感器还被应用到了医疗器械、高精密设备等领域。另外还开拓性地将视觉传感与触觉传感相结合形成主动感知系统，为机器人的发展开辟了新的方向。

在报告最后，孙富春总结了未来机器人在感知方面的研究框架，即主动感知、认知发育和行为决策要紧密结合，形成一个完美闭环。

以下为孙富春的演讲正文。

今天我要跟大家汇报关于机器人视触觉的传感器，特别是触觉传感器，包括机器人究竟需要什么样的感知。

一、机器人的发展

我们先来看下机器人究竟需要什么样的感知。机器人的概念是 1920 年捷克斯洛伐克的卡雷尔 . 开陪克的巨作《罗萨姆的万能机器人》中首次提出的概念。

其实机器人的发展与人类的发展密切关联，工具的发明只是将人的手延长，人仍然是动作行为的发出者。我们国家也有很多这方面的记载，如《列子》里面讲到西周时期能歌善舞的伶人，以及《古今注》中记载张衡做的机器人，它分为上下两层，且用到了世界上最早的减速器，所以世界上最早的减速器其实是中国人发明的。到了机器时代，由于蒸汽机的发明，机器实现了动作行为对人动作行为对人的实体分离，可以独立地从事很多工作。在计算机时代，机器人可以执行多任务的工作，而在网络时代，最大的改变是集群控制，即多个机器人协同工作。

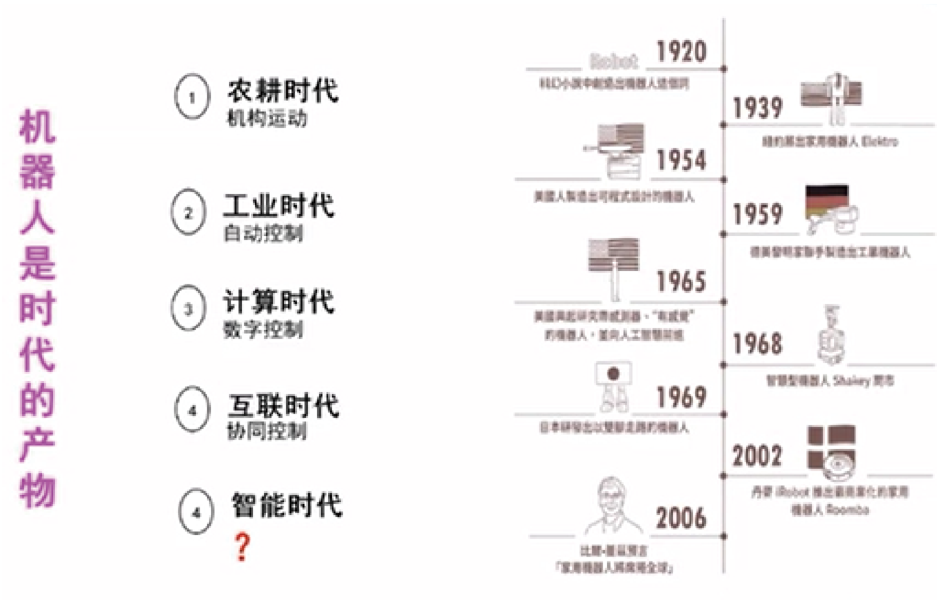


图 1：机器人的发展历程

让机器人能够代替人做非常精细化的工作是我们最大的梦想，其实让机器人拥有人类一样的行为，可能比交互和翻译更难。一方面，行为离不开脑袋越来越聪明，那么我们的人工智能算法能否做的更好；另一方面是脑体协同的问题，机器人的手要高自由度，且行为要与大脑关联在一起。

机器人想要学习人类，有三个问题需要解决：人工智能能否把神经中枢做好？机器人能否像人一样有感知？机器人能否拥有一双灵巧的手？这三者之间是相互关联的。

更高层次的，要想使机器人达到心灵手巧，首先要像人手一样高自由度，其次是能够进行分布式的控制，最后是要有需要聪明的大脑，使其能够学习和发育，面对变化无常的任务，包括复杂的产品，积累经验、形成知识。

二、视触觉传感器

下面介绍一下传感器的问题，我们国家在列出了 35 个卡脖子的问题，其中包括了光刻机、芯片技术、操作系统、航空发电等，其实还有触觉传感器。

图 2：35 项“卡脖子”技术和中国尚未掌控地 60 余项核心技术

当前的很多触觉传感器无法应用在实际环境中，主要原因是现在的触觉传感器分辨率不够高，无法比拟人手皮肤，特别是我们需要的三维触觉。

人工皮肤的研究主要分为两个方面：一部分是光电技术，比如通过电容、压阻以及光研究传感器；另一部分是研究人的皮肤表层的 Merkel 细胞，即通过生物传感的角度研究这种皮肤。

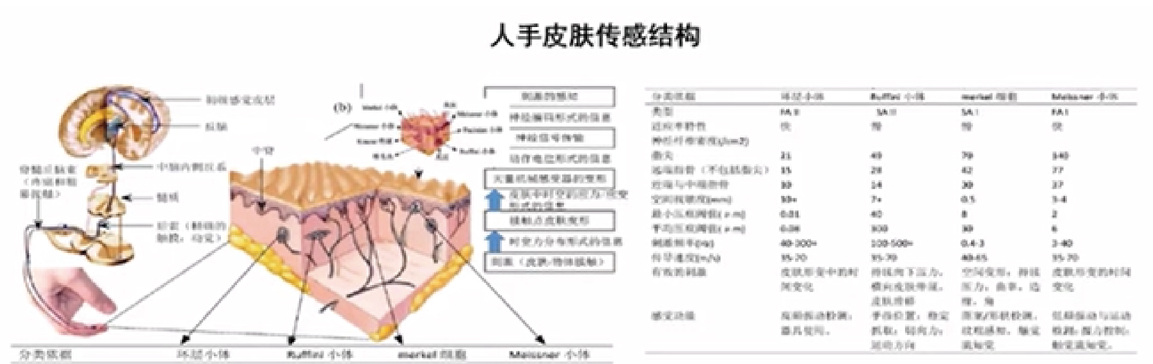


图 3：人手皮肤传感结构

我们团队现在研究的是通过微视觉实现指尖传感。目前可以实现空间分辨率在 1 毫米左右，密度 241 个毫米 / cm2，可以检测到 700Hz 的振动，可以感受到不同位置的两个接触事件的时间间隔为 30–50 毫秒，也可以检测到 40 μ m 的精细的表面纹理。

我们团队还提出了 1 个中心，再加上 24 个阵列的新型结构，它是模拟真实手指指肚的感知方式，通过通道占比分类器实现触觉的方向识别。

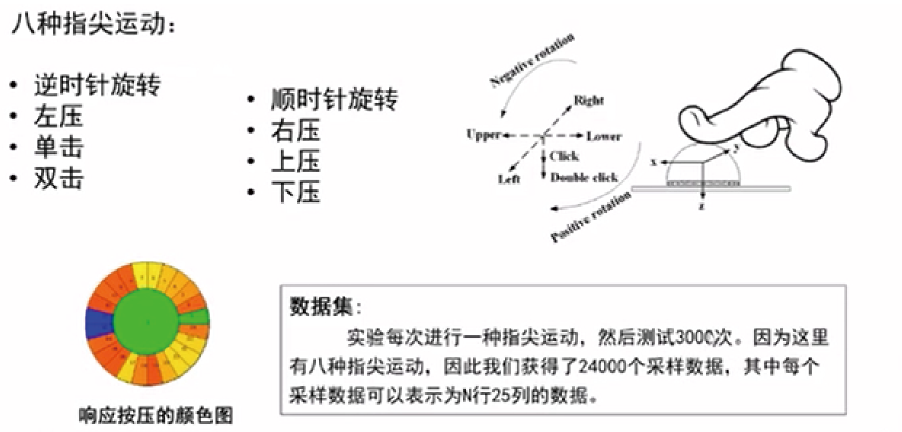


图 4：八种指尖运动

还有一种电容式传感器，我们做过一些滑动和振动试验，并与国际上近期做的一些工作做了比较，因为这项工作在国际上是属于最前列的。

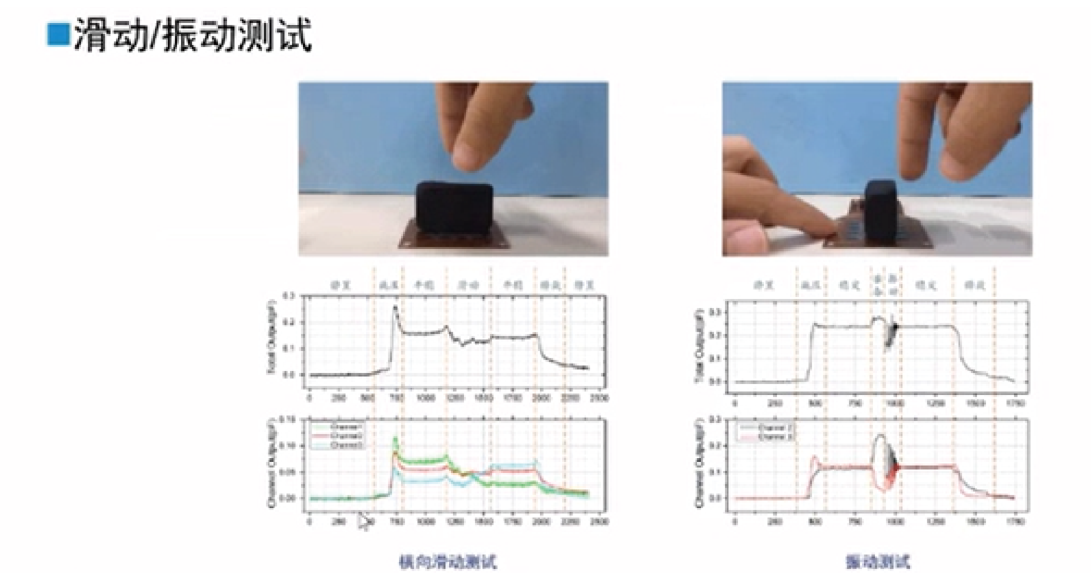


图 5：划动 / 振动测试

我们的第三代处理传感器在中医上进行了应用，主要目的是测量人的脉搏，并取得了得了较好的结果。我们也做了压阻式的处理传感器，包括抓取不同物体的触觉感知。我们的一款压阻型柔性触觉阵列传感器在横向滑动实验、抓取稳定性实验都获得了很好的表现，达到了产品级，并曾经在第七届电子信息博览会上展示过，它同样可以安装在桌面的小型机械臂上做为教学方面的实验器材。

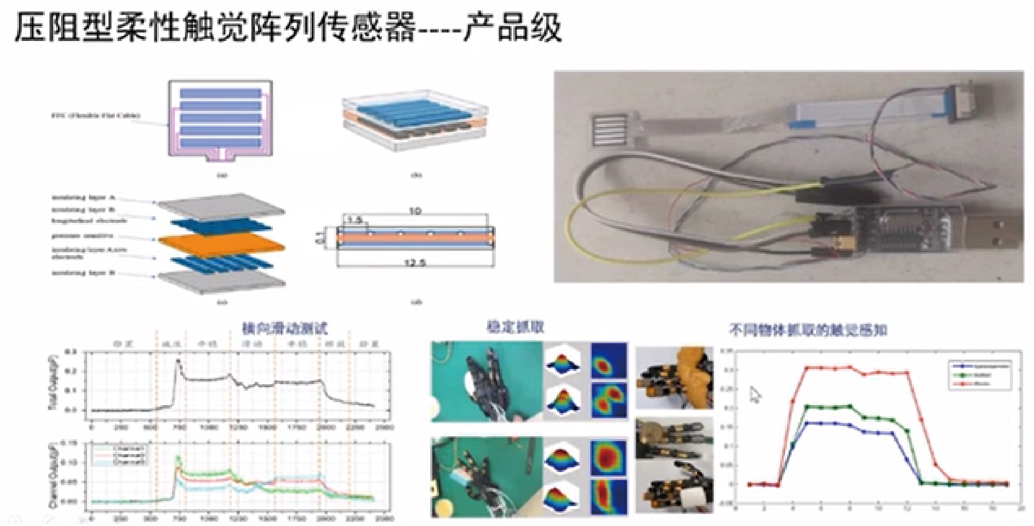


图 6：压阻型柔性触觉阵列传感器

结合仪器开发的触觉传感装置，我们建立了首个集成触觉信息和视觉信息的多模态主动操作公开数据集。此外，我们也开发了数据手套，研究人手感知操作特征，并建立了人手抓取操作的数据集，为机器人的精细操作提供了示数。

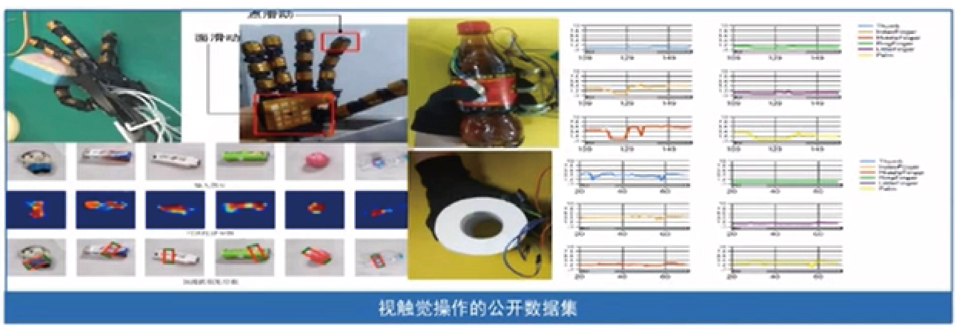


图 7：视触觉操作的公开数据集

图 8：基于数据手套建立的人手操作数据集

面向膝关节置换中压力测量问题，我们研制了医用的柔性处理传感器阵列，可以实时测量压力为成功膝关节置换手术提供保障。

微创手术过程中，存在因医生过力操作而导致的器官损伤问题，我们与山东省一家医院合作，研发了带有传感器结构的微创手术夹钳末端。

图 9：带有传感器结构的微创手术夹钳末端

目前我们正在做的、分辨率最好的一款指尖传感器，特点就是类似在人的指尖部分植入一个表层有大量的标志点的 LED 摄像机，如果手指在按动目标的时候产生形变，标志点也会产生相应的形变，摄像机可以将这个变形拍下来，接下来我们就要主要处理这个标志点移动的图像，进而获得物体表面的颜色信息、纹理信息，包括三维的触觉信息。目前我们也在开发温度信息，一种方法是将温感材料加在表层，能够很容易地测量温度，而且这种测量是不受环境结构光的影响，所以在结构光不好用的情况下同样能够测量温度。

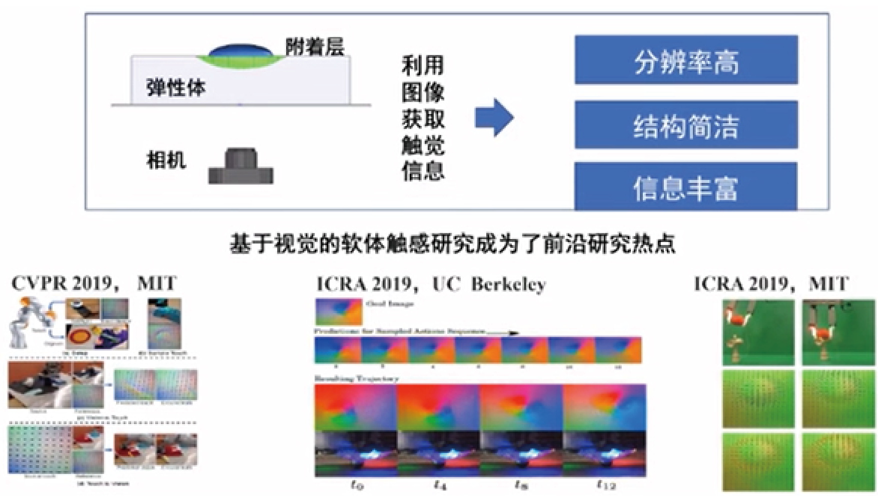


图 10：基于视觉的软体触感研究成为了前沿研究热点

通过视触觉传感器在相同部位采集的物体的光学信息和触觉信息，可以获得包含很多结构和深度的信息，可以用来做目标的识别和材质的识别，以及感知更深层次的纹理信息、温度信息等。

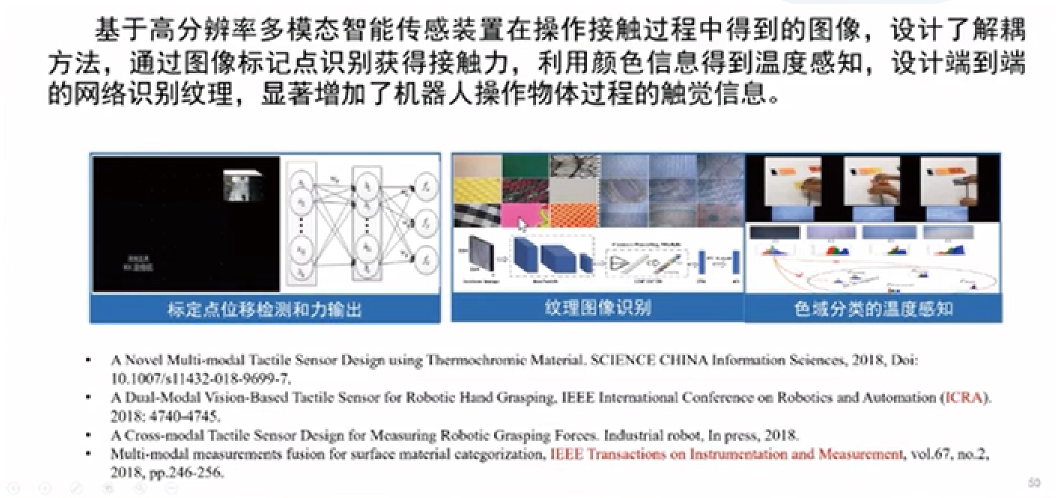


图 11：视触觉传感器显著增加了机器人在操作物体过程中的信息收集

我们将这个触觉传感器应用在机械臂的末端，使得非常细小的精细化操作都能够在其辅助下完成。多种实验结果表明，通过多模态的方法，可以有效识别其识别效率。

	触觉阵列	Thin-Film	Gelsight (MIT)	Gelforce	TacTip	TH-Tactile
原理	电容性	光学性	光学性	光学性	光学性	光学性
力	√	×	√	√	×	√
纹理	×	√	√	×	×	√
分辨率	高	高	高	中	低	高
温度	×	×	×	×	×	√
柔韧性	低	中	高	高	高	高
制造成本	高	高	低	中	低	低

图 12：高分辨率多模态智能传感装置

利用这套高分辨率多模态智能传感装置系统，可以进行认知方面的工作，即通过人的大量交互形成触觉的认知部分，并从局部感知扩展到全域感知。

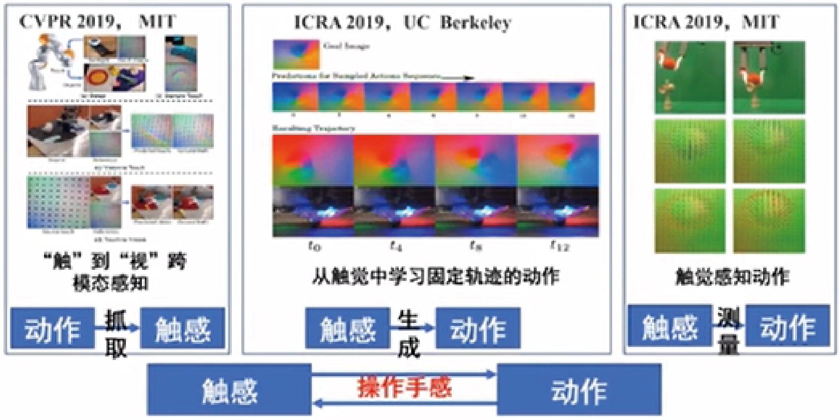


图 13：是否可以扩大机器人操作过程中的感知区域?

三、机器人的主动感知

以下是主动感知的概念：传统的感知有一个最大的缺点，即视觉系统仅仅根据传感器感受到的图像来做识别，但是人类是可以感知图像采集的好坏，并作出相应的调整。我们一方面利用传感器感知到的信息来做目标的跟踪、检测和识别，另一方面根据环境之间的交互，使得传感器感知的环境更加精细。二者结合起来进行行为分析，就是今天我们谈到的立体视觉部分，也是脑体协同部分。

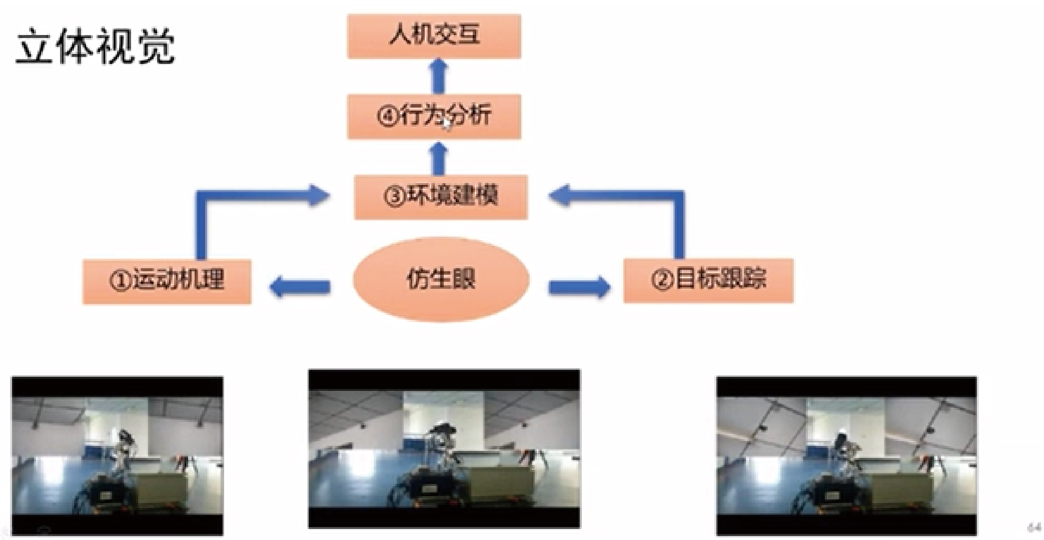


图 14：立体视觉

目前感知部分如果仅仅通过图片、通过机器学习、深度学习的改良是不够的，还需要通过脑体合作，不断去适应环境、克服传统视觉的脆弱性，通过交互不断提高鲁棒性和认知能力。

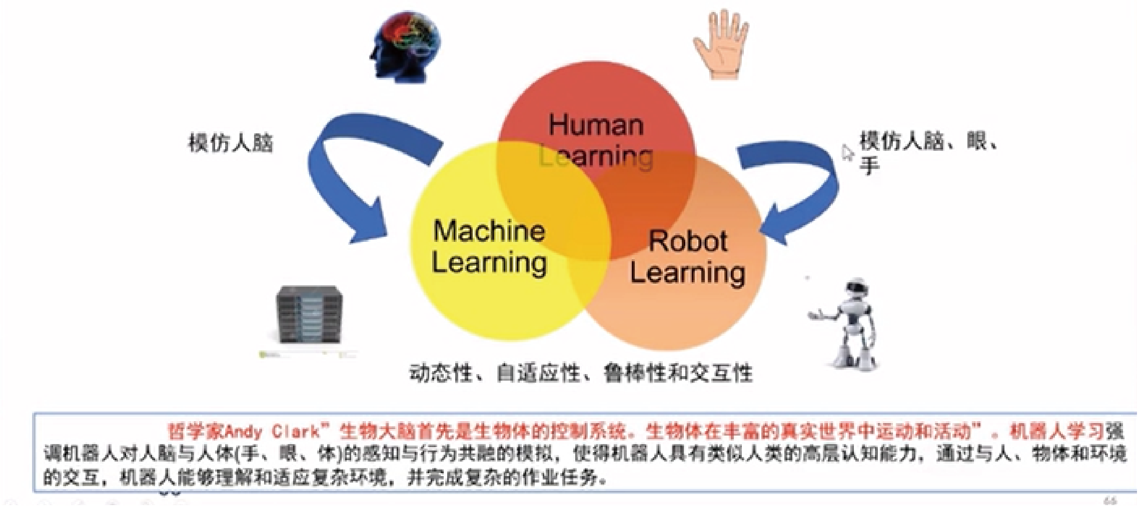


图 15：动态性、自适应性、鲁棒性和交互性

如何通过脑体结合提高机器人的感知能力？一是传感器要长腿，即主动感知； 二是要选择合适的传感器； 三是让机器人的传感器不断地增长知识和发育。

1）传感器要长腿，即主动感知

在这主动感知方面，我们的一个工作是通过特征金字塔网络和反向连接原理，增加底层语义，实现对小目标的识别。还有一个工作是通过组合的方法，结合深度学习方法，使其能够更好的检测以动物体，我们将这个工作应用到了咽试纸上面，获得了不错的表现。

触觉建模遇到的最大问题在于和时间相关，所以不能用一个静态的图像的方法来做触觉的建模。我们用线性动态系统的办法建立触觉模型，LDS 空间的数据在黎曼空间，我们采用核函数将其映射到欧几里得空间，通过这种方法很容易得到触觉的编码。这项工作也是 2016 年的一篇最佳论文。

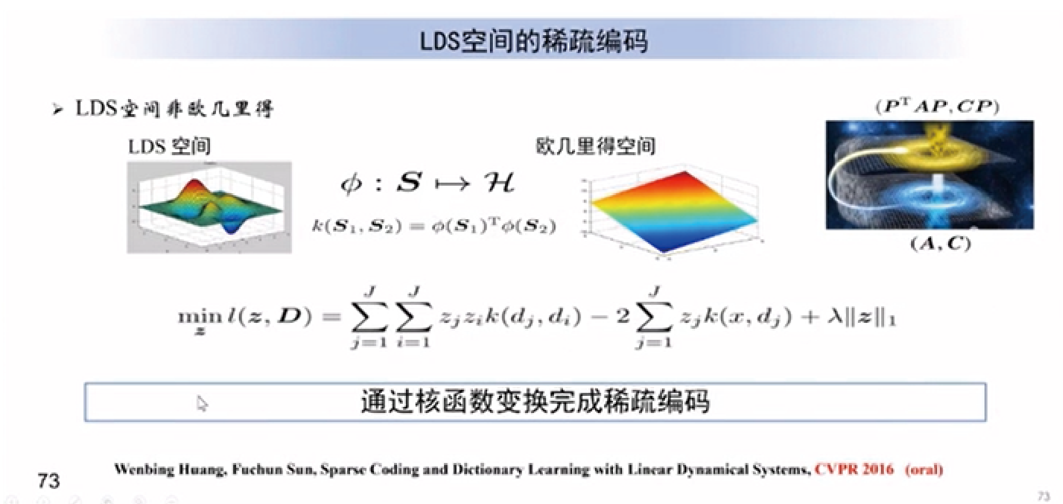


图 16：通过核函数变换完成稀疏编码

2）传感器的选择

机器人有很多传感器，包括视觉、听觉、触觉和雷达等，那么如何根据任务和场景进行选择？在这一方面，我们提出了主动感知网络，它能够根据需要选择传感器。我们还提出主动感知网络，学习模态间和模态内的特性，使其获得较好的感知能力。

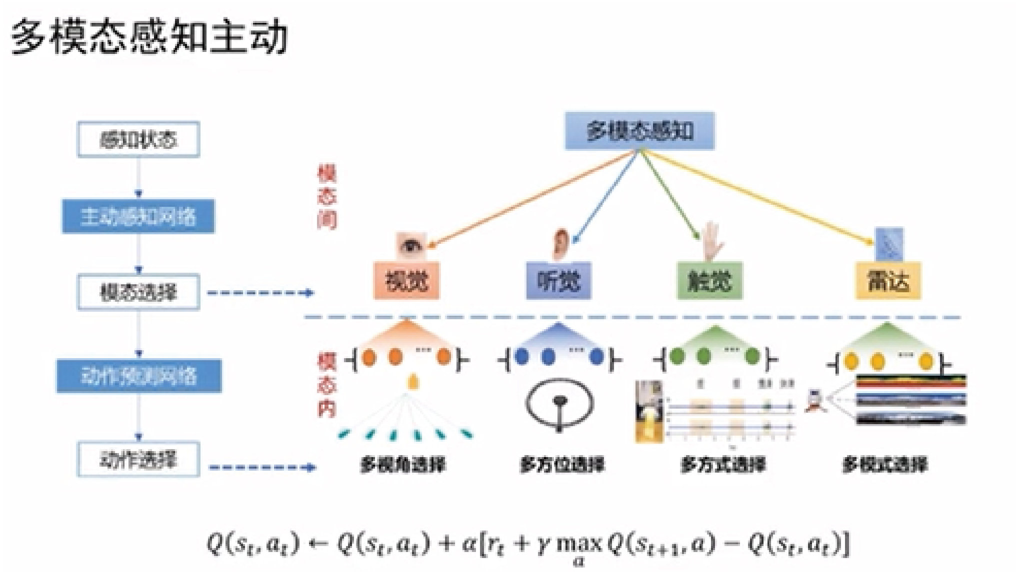


图 17：多模态感知主动

3）机器人传感器的增长和发育

发育的概念经常使用在教育中，但其实这个词是 1950 年图灵在谈到图灵测试是提及的。当前，我们团队也在结合触觉研究技能的学习发育，和感知的发育问题。

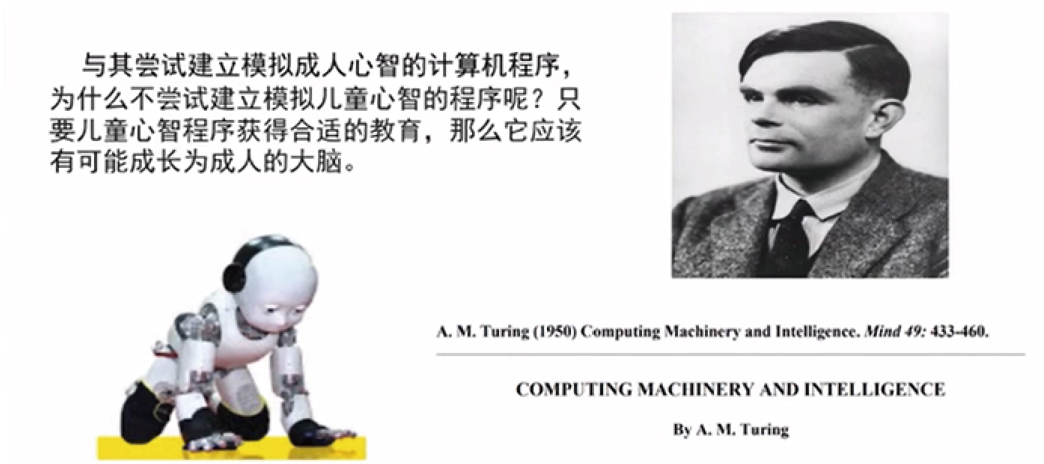


图 18：图灵提及的机器发育

四、面向感知的行为学习

这里主要讨论行为模仿和技能的增强，

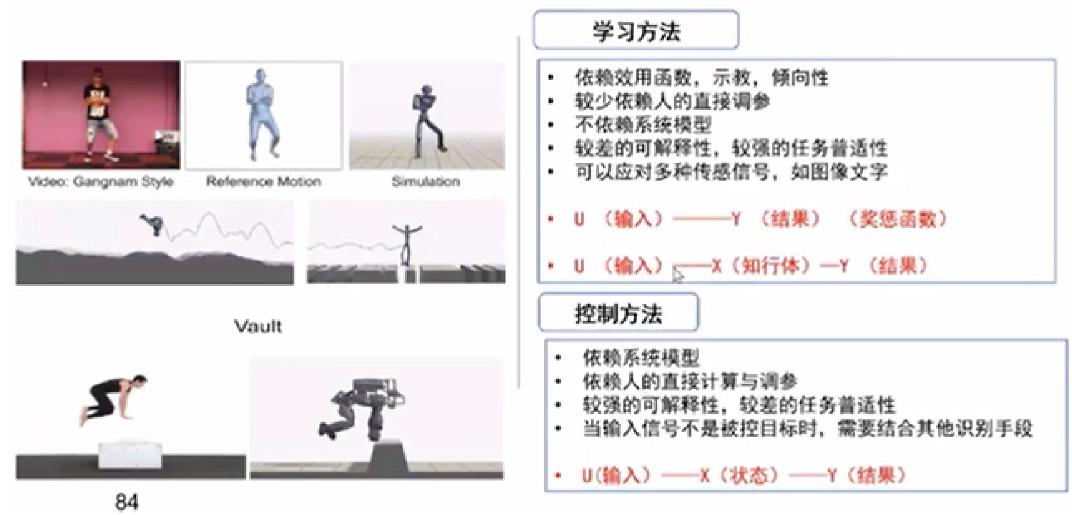


图 19：行为增强感知问题

这个问题主要包括两方面的工作： 一是如何面向观测不完备的情况？什么是观测不完备，如示教过程有手把手教和仅仅视觉示教两种方式，后者可能出现示教信息不完备的情况。

针对模仿和观测不完备的情况，我们在数学上严格证明了拖拽式技能模仿与观测式技能模仿之间缺少一种逆运动不一致的度量。利用这个不一致的度量能够实现技能的不断增强，我们也通过一个迷宫的实验验证了这一理论，结果表明我们的方法介于拖拽式模仿与纯观测式模仿之间。

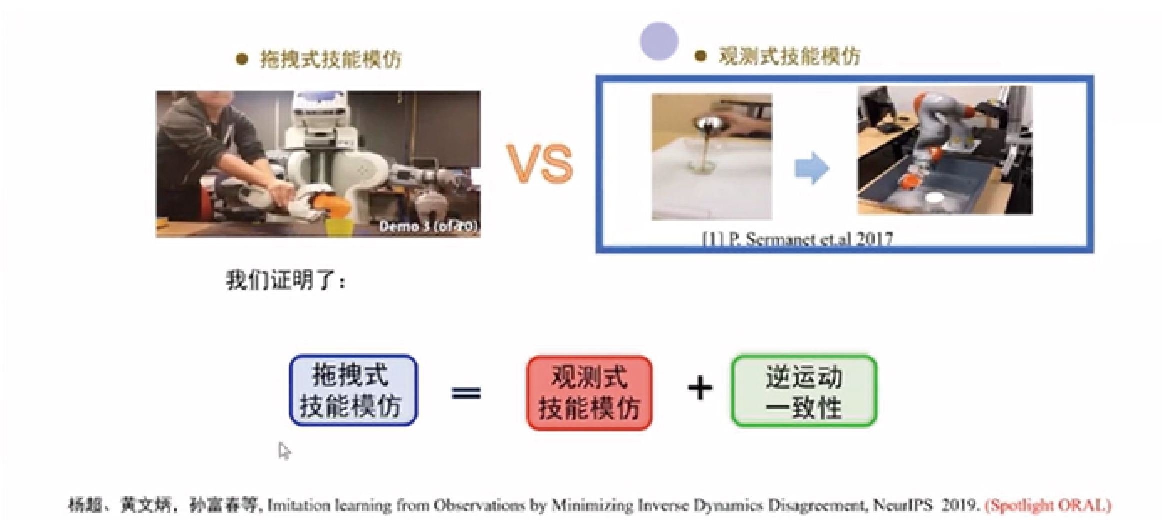


图 20：针对模仿和观测不完备情况的数学证明

针对示教信息不完备的情况，我们把基于示教的技能模仿成一个带约束的优化问题，通过带约束的优化，在专家示教附近寻找最优的技能策略。即把专家不断模仿的数据记录下来，构成一个数据空间，那么最优解一定是在该数据空间中，然后再通过强化学习进行约束优化，就能够得到比专家更好的工作。我们把这个工作应用在了弹钢琴，也获得了比较好的结果。

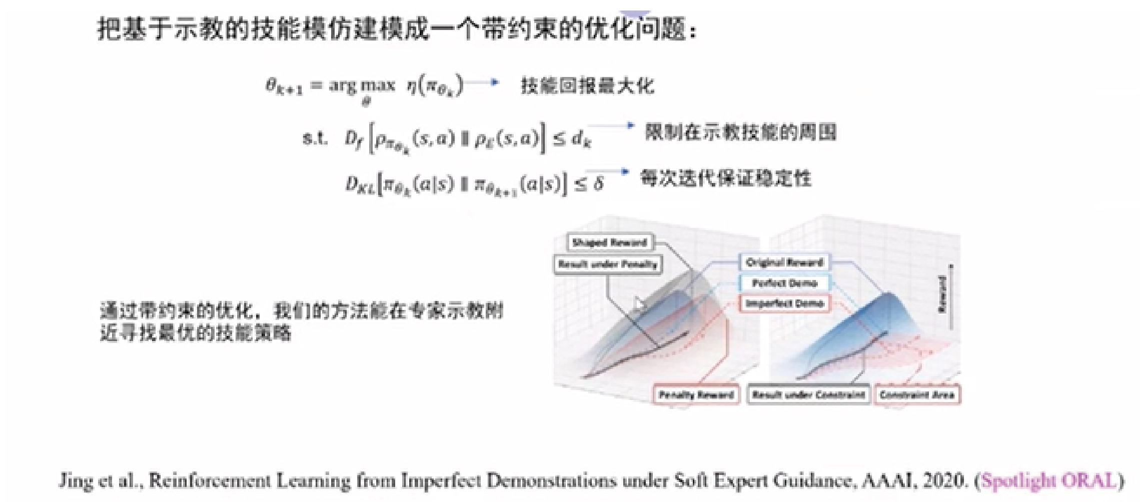


图 21：把基于示教的技能模仿建模成一个带约束的优化问题

五、未来展望

机器人究竟需要什么样的感知？机器人是不断与人和环境交往的，机器的学习一定是与行为、跟环境相结合的感知系统。所以我们需要能够面向动态交互式的环境和任务，行为的本质特性就是要动态、交互和适应。

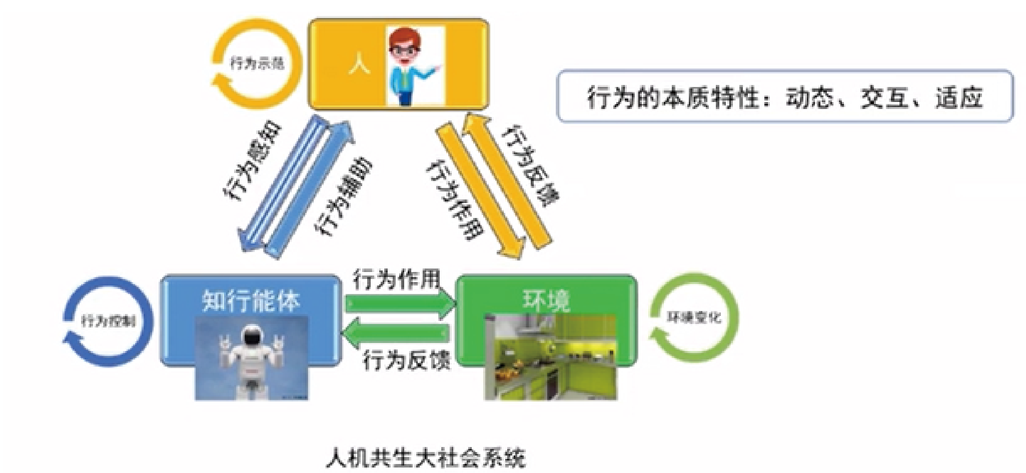


图 22：行为的本质特征

机器人的感知一定要主动面向动态适应和交互； 感知要发育； 要与行为决策时刻关联在一起，形成感知到大脑再到行为的闭环。

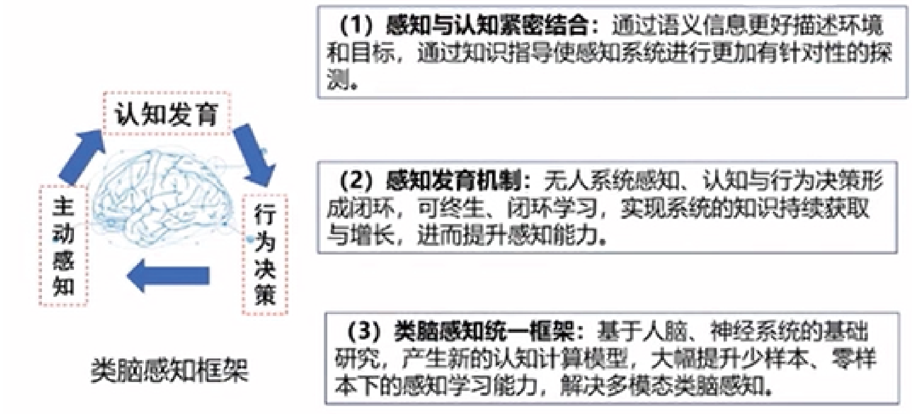


图 23：类脑感知框架

共融： 谈到目前机器人领域一个非常热的话题，就是共融的概念。是与作业环境、人和其他机器人自然交互、自主适应复杂动态环境并协同作业。很多人对共融特别感兴趣，包括很多戴假肢的人特别希望假肢能够像真的人手一样感受到疼痛，成为自己生活当中不可或缺的一部分。

协作机器人： 目前机器人还有一个重要特征是协作，将来要不断地与人打交道，怎样在这样的环境当中与人和谐相处，也是未来机器人需要面对的重要问题。

多模态感知实现反欺骗： 如何通过脑体协同实现这种反欺骗，是一个可以研究的方向。

联合感知： 随着云端技术、边缘计算的技术应用，我们尝试通过云端技术实现联合感知，包括不同机器人之间的感知、外场跨域的感知。

集群感知： 目前大量机器人在一起如何实现感知？因为机器人系统可以组成一个合成恐惧雷达，通过机器人系统形成一个半径，感知效果要比这些作为机械的加法要好，这也为未来人工智能领域提出了更大的挑战。

北大教授张大庆：基于 WiFi 和 4G/5G 的非接触无线感知 - 挑战、理论和应用

整理：智源社区 蒋宝尚

在第二届北京智源大会“机器感知论坛”上，北京大学计算机系教授张大庆做了题为《基于 WiFi 和 4G/5G 的非接触式无线感知：挑战、理论和应用》的报告。

张大庆，北京大学博雅讲席教授，欧洲科学院院士，IEEE Fellow，中国计算机学会（CCF）普适计算专委会主任。张大庆教授团队在国际上率先将 Fresnel Zone 模型引入到可无线感知领域，发现这是一个基于 WiFi 和 4G/5G 无线感知的新理论基础，其优点在于能够刻画收发设备位置、移动对象位置、速度与无线接收信号之间的关系，并能揭示各种感知参数对无线信号时域特征与频域特征的影响。

以下智源社区编辑是对张大庆报告的文字整理。

今天的演讲主题是基于 WiFi 和 4G/5G 的非接触式无线感知的挑战、理论和应用，整个报告会分为六个部分进行。

一、为什么选用 WiFi 和 4G/5G 信号做感知

图 1：WiFi 与 4G/5G 信号无处不在

WiFi 和 4G、5G 信号已经遍布在各个城市的各个地方，无论是在家还是办公室、酒店或者机场，人们都可以通过这些信号进行“连接”。其实，这里用的通讯介质主要是 4G、5G 信号或者 WiFi 信号的电磁波。

WiFi 和 4G、5G 信号除了能够完成通讯任务，还可以用它做感知的工作，这种感知手段和其他感知手段相比具有的优势是：1、无线信号无处不在；2、基于现存通讯设施，成本比非常低；3、无需穿戴感知装置；4、较小的隐私担忧。此外，相比摄像头，利用 WiFi 和 4G、5G 信号可以做穿墙感知。

有了上述优势之后，我们如何能够利用这些 WiFi 信号对人的行为状态进行非接触的感知？那么我来介绍一下无线非接触感知的基本原理和面对的挑战。

二、无线非接触感知的基本原理和面对的挑战

WiFi、4G 和 5G 信号是利用无线电磁波来做感知，而无线电磁波在空间的传播实际上是沿着多径传播，另外还需要明确的是环境中静态物体反射、衍射带来的静态路径信号基本是不变的，但当环境中有人或物体活动时，会带来传播路径的连续变化。因此，我们可以通过分析描述传输路径变化的通道信息 CSI，来反推和感知环境与人的情境。换句话说，能够用 WiFi CSI 刻画路径信息。

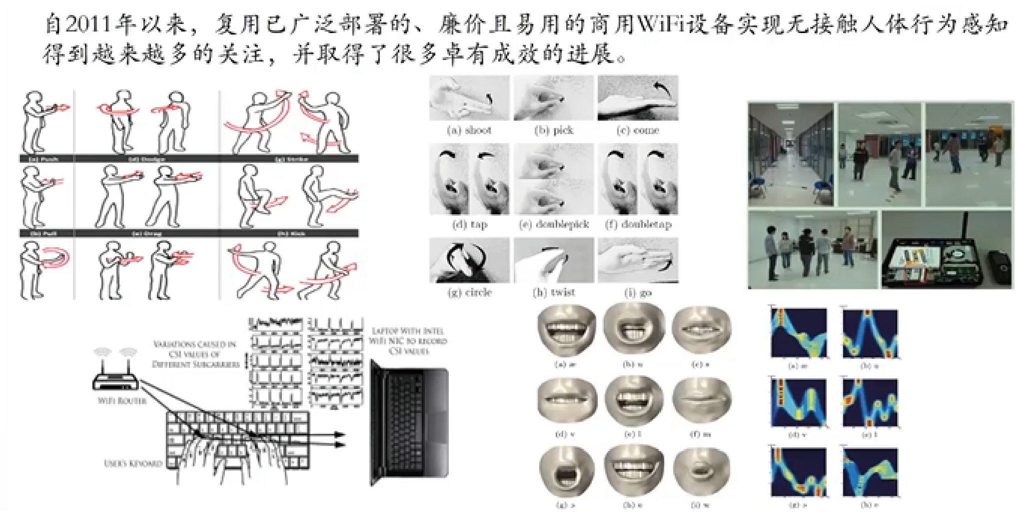


图 2：基于 WiFi CSI 的人体行为感知应用

自 2011 年以来，复用已广泛部署的、廉价且易用的商用 WiFi 设备实现无接触人体行为感知得到越来越多的关注，并取得了很多卓有成效的进展。国内各个院校也做了大量有意思的工作，比如南京大学的键盘输入项目，香港科技大学的唇语识别项目，清华大学的跳舞步伐识别的项目，西安交大的人数清点项目等等。

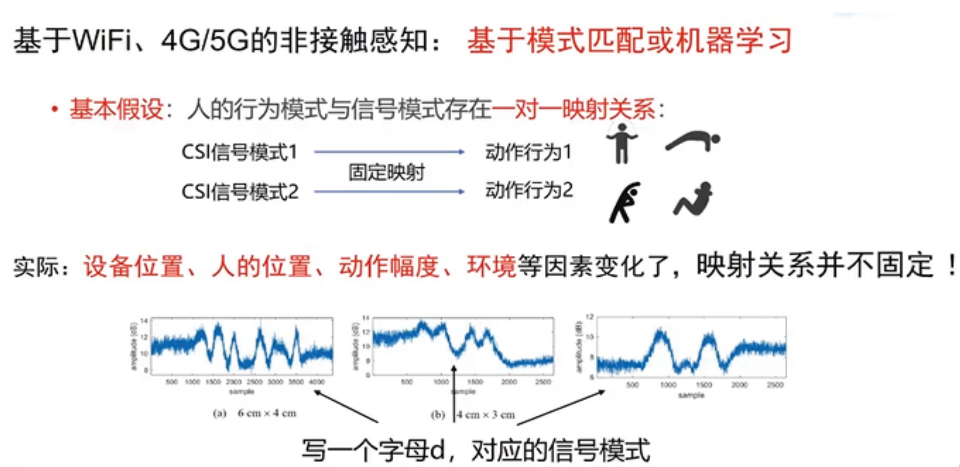


图 4：众多识别项目的基本假设

纵观这些识别工作，大多都有一个共同的假设，就是当人做某一个特定行为的时候对应的 CSI 信号模式是几乎恒定的，也就是说，人的行为模式与信号模式存在一一映射关系。而在实际中，设备位置、人的位置、动作幅度、环境等因素变化了，映射关系并不固定。因此利用这个假设进行的模式识别和机器学习，到目前为止性能都不稳定。如上图所示，书写字母 d 的手势对应的信号模式是不一样的。

1. 基于模式和学习，采样、标注、选参不易，信号模式易变，很难到100%的识别率

2. 缺乏对背后感知机理的理解，WiFi能感知什么不能感知什么？信号的变化模式与人、WiFi设备和环境之间有无一个定量的映射关系？

对两个重要的问题缺少答案

1. WiFi感知的理论基础是什么？
2. WiFi感知的边界在哪里？

图 5：WiFi 感知领域的共性挑战

另外，在过去几年的研究里，我们发现已有 WiFi 感知领域，大部分采用的方法都是基于模式识别、机器学习，因此会存在两个重大的共性挑战： 第一，基于机器学习和模式识别的方法需要大量采样，也需要标注，特别是环境、行为和人的位置发生变化时，信号的模式会发生变化，很难做到稳定准确的识别率； 第二，缺乏对背后感知机理的理解，WiFi 能感知什么不能感知什么？信号的变化模式与人、WiFi 设备和环境之间有无一个定量的映射关系？现在这些问题都没有答案。

其实，归根结底是对两个重大科学问题没有答案：WiFi 感知的理论基础究竟是什么？ WiFi 感知的边界在哪里？

1. 提出一套通用的无线感知模型、理论和技术
2. 揭示无线感知的机理和感知极限
3. 在真实居家环境构建一套实时、连续的监测、识别系统

图 6：3 个研究目标

针对上述两个重大科学问题，我们团队过去五年，主要是针对三个问题进行研究：1、试图提出一套通用的无线感知的模型理论和一系列相应的技术；2、通过模型理论试图揭示无线感知的机理和感知极限；3、在上述理论的基础上，希望在真实居家环境当中构建一套实时的、连续的检测、识别系统。

三、Fresnel Zone 模型：一个基于 WiFi 和 4G/5G 无线感知的新理论基础

具体而言，为了研究上述问题，我们将 Fresnel Zone 模型引入到无线感知领域，发现这可以成为一个基于 WiFi 和 4G/5G 无线感知的新理论基础。

下面详细介绍一下 Fresnel Zone 模型，以及基于此模型电磁波传播的重要特性，从而探讨一下 WiFi 感知的极限到底在哪里。

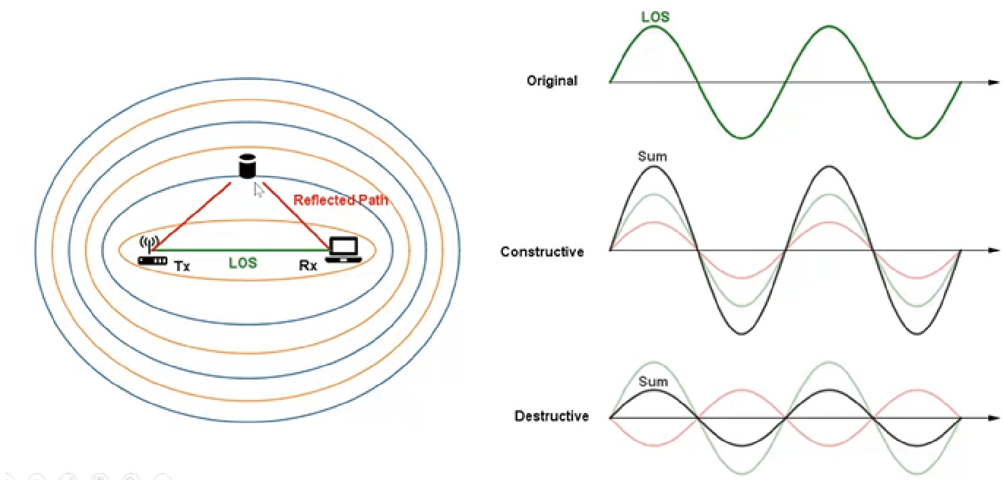


图 7：空间中的静态物体如何影响信息接收（一）

假定在空间当中有一对收发设备，把发送端和接收端作为椭圆焦点，通过这两个椭圆焦点，如果让反射路径减去只是路径的长度为半波长的整数倍，能够画出多个椭圆，我们把这一系列椭圆定义为 Fresnel Zone。

那么空间中一个静态物体如何影响接收信号？如上图所示，TX、RX 是 WiFi 的收发设备，以 TX、RX 为焦点，“辐射”出了多个 Fresnel Zone，假定环境里有一个静态物体，处于 1st Fresnel Zone 位置，让 TX（发送端）发送一个正弦信号，如此在接收端就会形成两条路径，其中，反射路径比直射路径（直径）长 1/2 波长，因此就会有π的相位偏转，如果加上反射带来的相位偏转，就会形成一个相位相同的叠加类信号。

如果物体位于 2nd Fresnel Zone，这时候形成的反射路径比直径长一个波长，由于一个波长对应 2 π的相位偏转，在加上反射本身带来的 180 度相位偏转，就会呈现出“反向”的相位，如此反射路径和直射路径 LOS 就产生了彼此相消的效果。

通过上述分析，我们可以得出静态物体如何影响接收端信号的结论：当物体在序号为奇数的 Fresnel Zone 时会产生增强的信号，当物体在序号为偶数的 Fresnel Zone 时会出现信号削弱的情况。

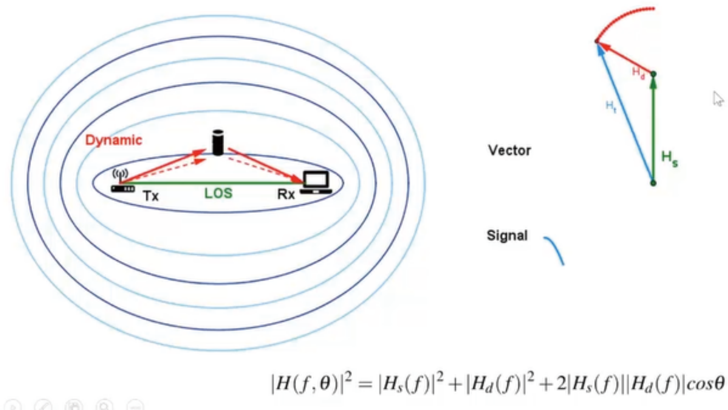


图 8：空间中的静态物体如何影响信息接收（二）

上面是静态物体的情况，那么动态物体如何影响接收信号呢？首先把静态路径的信号定义为 Hs，把动态物体的反射路径定义成 Hd。因此当物体在动的时候，接收信号相当于 Hs 和 Hd 两个向量的叠加。另外，动态向量的相位是慢慢发生偏转的。所以，动态向量每变化一个波长，那么它就会围绕着 Hs 旋转一周（360 度）。如此便能在接收端得到一个波峰、波谷交替出现的类似正弦波的信号，物体穿越 Fresnel Zone 的边界线（例如 1st 和 2nd 区域中间的那条线）正好对应波峰，或者波谷。

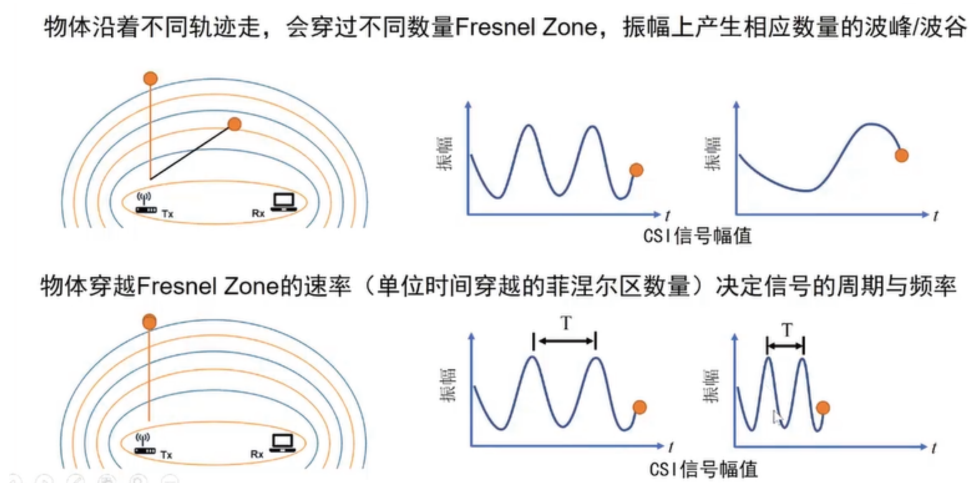


图 9：Fresnel Zone 与无线信号的时域与频域特征

介绍完 Fresnel Zone 模型的原理、性质，我们来讨论一下物体的活动与无线信号的时域与频域特征的一些关系。如上图所示，当物体沿着不同轨迹行走时，会穿过不同数量的 Fresnel Zone，振幅上产生相应数量的波峰 / 波谷； 物体穿过 Fresnel Zone 的速率（单位时间穿越的 Fresnel Zone 数量）决定信号的周期与频率。因此物体移动的很多参数都会影响到无线信号的时域与频域一些信号特征。

1. Fresnel zones → 一组同心椭球体
2. 幅值的波峰和波谷 → Fresnel zones 边界，信号周期对应相位变化2π的时间，信号频率对应单位时间切割的菲涅尔区的数量
3. 路径长度变化 ≥λ (e.g., 5.7cm for 5.24GHz)
   - 相位变化 ≥2π → 产生的信号 呈现类似正弦的一个波形
4. 路径长度变化＜λ
   - 相位变化 ＜2π → 产生的信号 是类正弦波形的一个片段

图 10：WiFi 电磁波信号传播的一些性质

小结一下 WiFi 电磁波信号传播的一些性质：1、Fresnel Zones 对应的是一组同心椭球体；2、幅值的波峰和波谷和 Fresnel Zones 边界有关，具体而言，信号周期对应相位变化 2 π的时间，信号频率对应单位时间切割的 Fresnel Zones 的数量；3、当路径每变换一个波长，相位变化是 2 π，产生的信号呈现类似正弦的一个波形；4、当物体稍微有些移动时，路径长度变化小于一个波长，相位变化小于 2 π，产生的信号是正弦波形的一个片段。

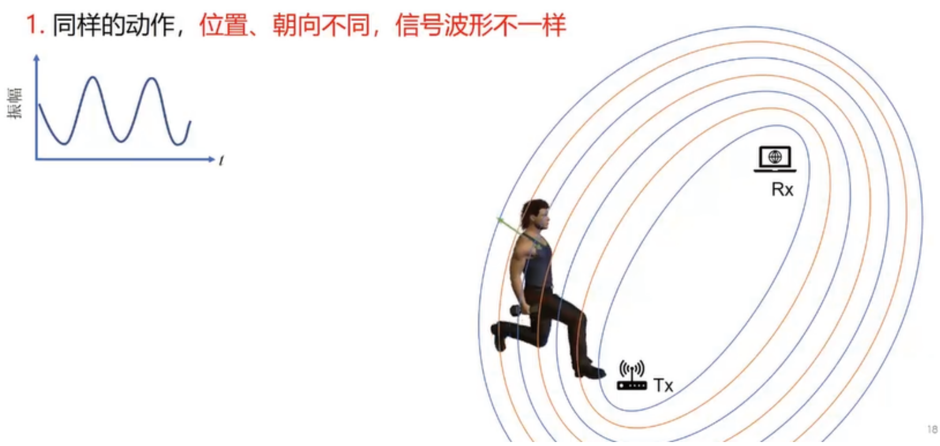


图 11：粗粒度动作对应信号波形与位置、朝向的关系

介绍完性质，下面来看看当人在做不同动作时候，信号的波形与位置、朝向有何关系。我们先让一个人做同样的动作，在不同的位置，会得到类似上图左上角波型，换一个位置再做同样的动作，会得到一个不同的波型；如果让一个人在同一个位置进行速度不同的动作时，信号频率也会不一样。

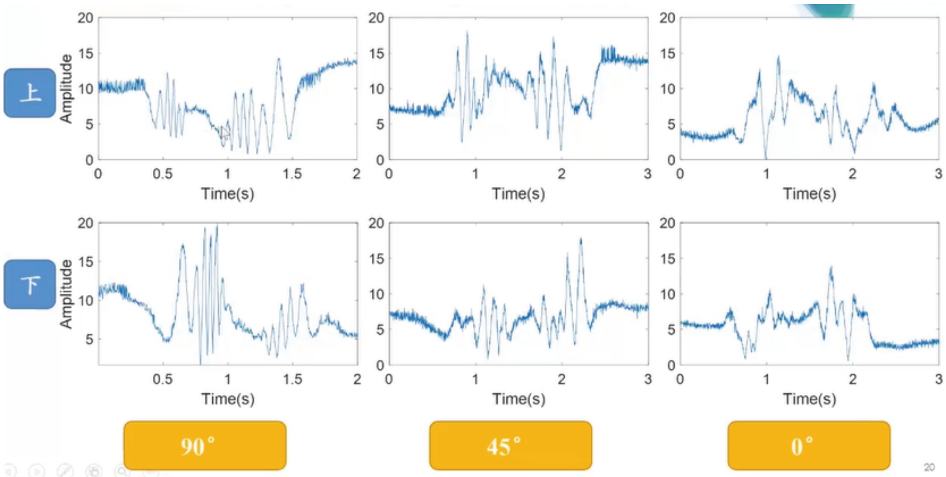


图 12：三个朝向的“上 / 下”手势动作 vs 接收信号模式的变化

上面是理论的估算，具体实验结果如何呢？因此我们设计了一个实验，让人做几个简单的动作：挥动手臂上下或左右运动，在不同的方向、不同的位置，得出的结论如上图所示：不同的位置，接收信号模式是不同的。

当然，上面是大幅度动作的实验，我们也进行了小幅度的动作实验，得出的结论是：同样的动作，Fresnel Zones 位置、朝向不同，信号波形也不一样。

对 Fresnel Zones 模型性质推论进行小结一下：1、在不同位置、朝不同方向，同一行为会产生不一致的信号变化模式；2、在不同位置，做不同行为，可能产生类似的信号变化模式；3、仅基于少量采样和机器学习，无法实现稳定的行为识别。

对5GHz频段的WiFi，其波长λ在5.7cm左右。根据理论模型，当目标的位移导致的反射路径长度变化5.7cm时，动态向量刚好旋转一周，产生一个完整的正弦信号；即路径长度变化Δd与向量旋转角度ϑ满足下面公式：

$$\frac{2\pi}{\lambda} = \frac{\vartheta}{\Delta d}$$

例如：5mm左右的体动带来的反射路径长度Δd变化1cm时，向量旋转角度ϑ大约是1.1弧度（63度）。而1毫米的人体位移对应相位旋转在12.6度。那么，Wi-Fi是否具备毫米级的感知能力，关键在于是否足够检测到这不足12.6度的信号片段。这即是WiFi感知极限评判的理论基础。

27

图 13：WiFi 信号感知的理论极限

下面用 Fresnel Zones 的模型对 WiFi 信号感知的理论极限进行估算。对于 5GHz 频段的 WiFi，其波长在 5.7cm 左右。根据理论模型，当目标的位移导致的反射路径长度变化 5.7cm 时，动态向量刚好旋转一周，产生一个完整的正弦信号；因此，路径长度变化与向量旋转角度满足公式如上图所示。

例如；5mm 左右的体动带来的反射路径长度变化 1cm 时，向量旋转角度大约是 1.1 弧度（63 度）。而 1 毫米的人体位移对应相位旋转在 12.6 度。那么，WiFi 是否具备毫米级的感知能力，关键在于是否足够检测到这不足 12.6 度的信号片段。这即是 WiFi 感知极限评判的理论基础。在实验层面上，我们发现 WiFi 信号的感知极限是毫米，可以通过 WiFi 检测到人的呼吸，而心跳比较难检测到。

四、基于 Fresnel Zones 模型的感知应用举例

前面讲到无线信号传播的有关性质，给出了 WiFi 感知行为的极限，下面我们利用这些简单的性质和模型实现一些具体的应用，包括呼吸监测和手势识别。

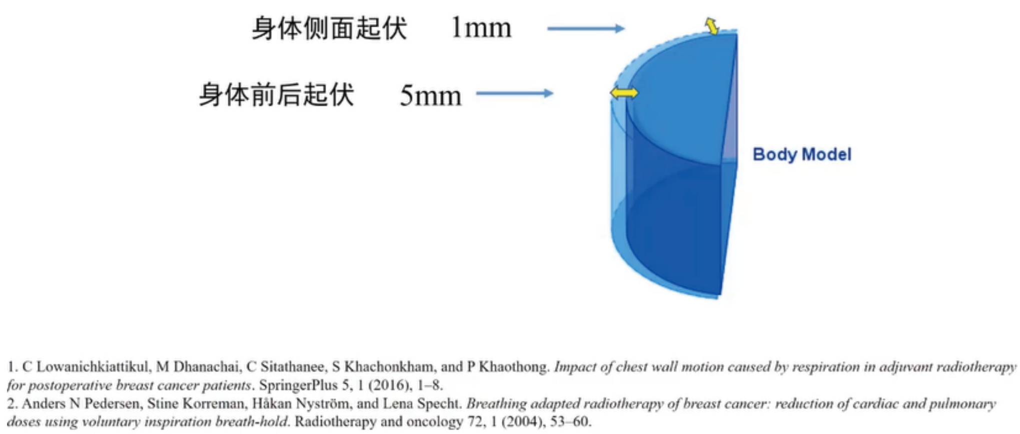


图 14：呼吸检测 – 人体建模

为了利用 WiFi 信号对人的呼吸进行监测，我们首先对人体进行简单的建模：把人看成一个半圆柱体。呼吸的时候前胸大概会有 5 毫米的位移，侧面大概有 1 毫米的位移。因此，用 WiFi 信号感知人的呼吸的问题，相当于要用 WiFi 信号感知到 1 毫米的人的体动。

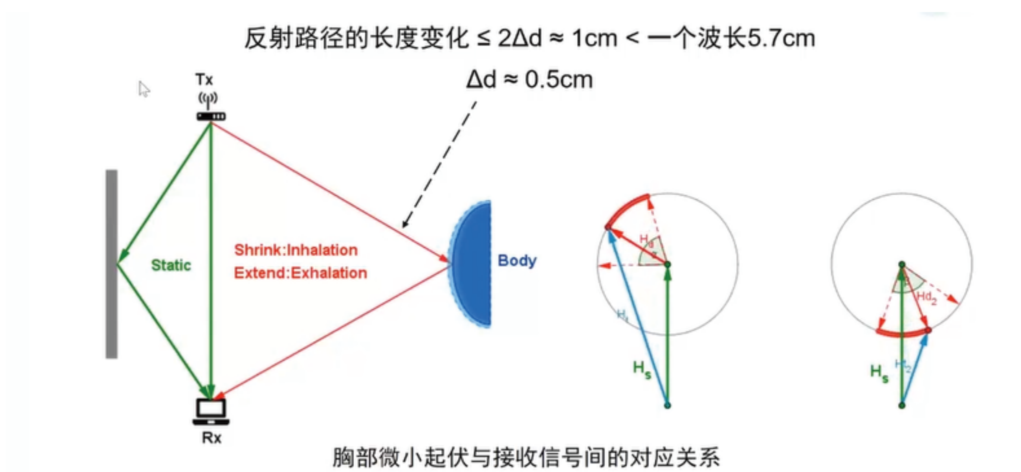


图 15：人体呼吸引起的信号变化是 Sine 周期的一个片段

具体建模分析图示如上。5 毫米位移，根据中学的几何知识，红线（反射路径）大概会变化 1 厘米，而一个波长是 5.7 厘米，所以 1 厘米对应的大概是 60 度的相位变化，至于对应哪个 60 度的片断完全取决于相对于收发设备的位置。

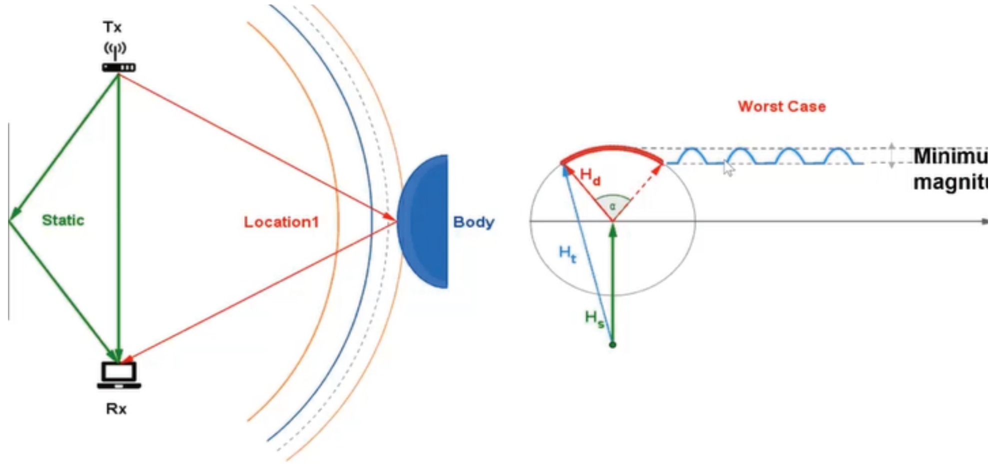


图 16：观察一：最好 / 最坏位置 –>Fresnel Zone 中部 / 边界

考虑两种情况：1. 假定人的前胸刚好就在 Fresnel Zone 的边界呼吸，一呼一吸刚好对应最上面 60 度的相位变化，那么振幅对应微小的波动，其频率刚好等于人呼吸频率的 2 倍，所以这种振幅的微小变化很容易被噪声淹没，人的呼吸比较难以监测。但是，要让人稍微往前移动 1~2 厘米，处于两个 Fresnel Zone 边界线的中间，那么呼吸可能带来比较大的振幅变化，由此可以看出，如果人在菲涅尔区中间的时候，信号本身振幅变化的波动很容易被监测得到。

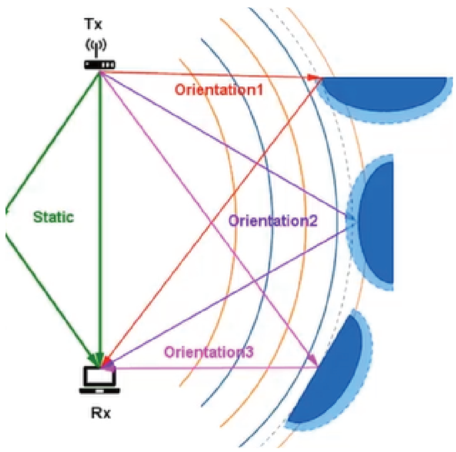


图 17：观察二：人体朝向 –> 有效位移

2. 人的朝向。针对上图三种情况能够有三种朝向位置，得出的结论是：正对时可以很好地监测到人的呼吸，背对时很难检测。根据这个原理，我们能够画出呼吸能否被监测到的热力图，可以看到“被监测”和“不可被监测的区域之间是交替出现的，可能存在一些所谓的盲区。在盲区中，呼吸靠振幅是比较难监测的，因此，为了

解决呼吸的盲区监测和信号模式不稳定的问题，我们团队提出利用振幅与相位信息构建正交感知信号，从而解决盲区和信号不稳定的问题。

构建正交信号的观察是：当相位不好时，振幅“表现”优秀，当振幅表现不好时，相位表现优秀。因此通过这种互补的关系，总是能够通过合适的投影准确监测人的呼吸。

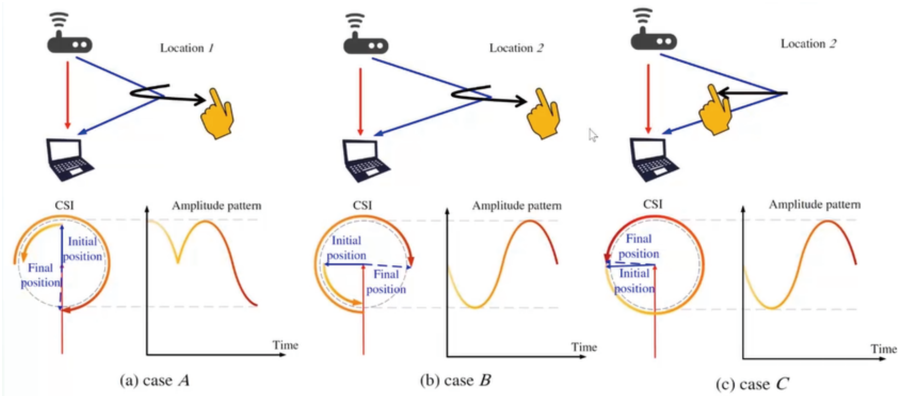


图 18：应用实例 2：手指动作识别

在手指动作识别实例中，我们让人用手指在空中简单画一个弧。可以看出手指头在动的起始位置不一样，得到振幅的变化波形便不一样。然而，如果我们让一个手做一个单一类的动作，然后会发现得到的信号变化和上图(b) 反而是一样的。因此，我们如果简单地利用振幅信息进行识别的话，可能会发生识别错误。

如何解决这个问题呢？我们可以通过信号变化技术实现，首先我们得到一对正交的振幅和相位信息，然后通过观察不同手势之间带来的相位变化，可以发现两个动作一样（位置不一样）的时候相位也是一样的；当另外一个动作不一样（位置一样，动作不一样）的时候相位变化也是不一样的。因此，可以通过信号变换的技术做到对手势活动的准确识别。

五、其他无线感知应用于视频演示

图 19：其他 WiFi 感知相关应用

前面介绍了两个具体的实例，下面就给大家展示我们团队最近几年已经开发的一系列 WiFi 应用。包括跌倒监

测、室内行走方向的监测、轨迹的跟踪、室内的定位以及呼吸和睡眠的监测等等。

六、结论

现在对整个报告做一个总结:

1. 基于 WiFi 等无线信号的非接触感知是人体感知的一种理想方法;
2. Fresnel Zone 模型是无线感知的一个新的理论基础。它刻画了收发设备位置、移动对象位置、速度与无线接收信号之间的关系，揭示了各种感知参数对无线信号时域特征与频域特征的影响。
3. Fresnel Zone 模型告诉我们: 在不同位置、朝不同方向做同一行为会产生不一致的信号变化模式; 靠数据采集很难穷尽各种场景和可能，不了解机理而简单利用机器学习无法做到稳定的、100% 的识别率。
4. 只有理解理论模型，通过信号变化并选定与目标位置、朝向、动作幅度无关的信号特征，才能构建出稳定可靠的无线行为识别系统。

问答

观众：通过 WiFi 的方式确实保护了隐私，是不是也有安全方面的挑战？比如可以直接侦测 ATM 机取款密码？

张大庆：很明显，任何一种“感知”都存在矛和盾的辩证，无线信号感知能够完成识别任务，但在某种意义上也会泄露人的隐私。矛和盾的关系始终是存在的，但是无线信号还是有它的优势，我们在和平环境当中用信号感知人的行为可以帮助人。但也需指出，如果感知过程中，有人有意地干扰信号，那么无线信号的行为识别工作就无法完成。

观众：手势识别的精准度怎么样?

张大庆：根据我们的设计，识别率可以达到 96%–97%。

观众：如果房间当中有多个人的话，呼吸的检测会不会更复杂一些?

张大庆：是复杂一些，今年我们正好做了这方面新的工作，在一张床上四个人都没有问题。

观众：最近 MIT 有一个报道跟你的应用类似，可以检测新冠患者居家呼吸活动，是不是真的？距离市场化有多远?

张大庆：MIT 的工作我们很熟悉，他们团队当中主要的几个人都到我们北大访问过，并且最主要的学生也是我们北大过去的，但他们用的是连续波雷达做的，不是普通的家用 WiFi，他们做的效果是相当不错的。

观众：能不能对人的身份进行识别?

张大庆：有些团队的同事已经做过一些工作，通过识别人的步态可以对人的身份进行识别，但是究竟能够识别到多少人可能还很难说。

观众：WiFi 的多径效应对识别有没有什么影响和作用?

张大庆： 这个多径是客观存在的现象，一般来说都会把 WiFi 多径当成有害的东西，实际上我们发现多径可以被很好地利用，利用这个现象可以做很多有意思的事情，包括对微小行为的识别也是利用了多径现象。

观众：无线信号天线相除的应用是什么?

张大庆：普通的 100 多平米的家居当中有一对设备的话就可以把室内人的呼吸检测出来。

观众： 无线感知受环境影响比较严重，请问解决环境依赖问题有没有什么好的方案？无线感知要想真正大规模推广还要做哪些努力?

张大庆： 我们提出的 Fresnel Zone 模型揭示了无线感知依赖于环境、依赖于位置朝向等因素。通过此模型引入可以定量地刻画信号变化与环境、人、位置和朝向的关系，有了这个指导以后再去做就比较有针对性，至于该如何做不依赖于环境的行为识别仍然是一个难题，最近我们团队在这方面也有一些新的进展，今年的文章也有几篇是关于这个问题。

观众： 您现在的工作是一收一发的情况，还是也包含一发多收或者多发多收？有没有在物理的理论模型和机器学习互相结合的方面的工作?

张大庆： 实际上我们现在用的大部分都是一发多收，可以推广到多发多收，一发多收是比较常用的，家里一般有一个 WiFi 的路由器，接收装置可以很多、冰箱、彩电、空调，我们可以利用这些装置对家里比较密集地进行感知。目前我们做的很多行为识别工作都是把物理模型和机器学习深度结合，所以我们做到 96%–97%，通过物理模型把信号本身做了比较好的变换，利用变换然后再利用深度学习和机器学习来做，效果才能达到比较理想的情况。

观众：是不是要用什么特殊的 WiFi ?

张大庆： 正常的 WiFi 就行，商用 WiFi 的信号都能满足，只是厂家有的时候没有把这个信号让大家用 API 的方式存取而已。

嘉宾：WiFi 信号和雷达信号有没有区别？这套理论能不能用于雷达?

张大庆： 我们知道 WiFi 信号和雷达信号肯定是不一样的，设计的目标也是不一样的，因此我们无线感知领域当中有人用雷达来做感知，有人用 WiFi、4G 和 5G 来做感知，但是要做公平比较的话，我们发现雷达的能力会更强，因为雷达的带宽更宽，也是专用设备，专门为感知设计的，WiFi 主要是为通讯设计的，因此感知的辨识度和精度都没有雷达那么好。我们考虑因为 WiFi4G 和 5G 信号 Cost 很低，居家环境也可以做些比较普适的功能。至于原理上有些是可以被用在雷达里面，有些可能是比较适用于 WiFi 或者 4G 和 5G 信号。

清华教授史元春：人机交互从精准走向模糊

整理：智源社区 沈磊贤

在第二届北京智源大会“机器感知”专题论坛中，清华大学计算机系史元春教授做了题为《人机交互从精准走向模糊》的报告。

史元春，清华大学全球创新学院院长，“长江学者”特聘教授。其科研成果近年连续获得国际人机交互领域顶级会议最佳论文奖，并两次获得国家科技进步奖。她的主要研究方向为人机交互、普适计算、多媒体、网络教育技术等。在报告中史元春阐释了触屏、VR 等自然用户界面交互效率降低背后的科学问题，并介绍基于手指运动控制能力贝叶斯模型的智能文本输入方法、基于交互行为时序模型的动作意图准确判别方法等最新研究成果。

一、从 GUI 到 NUI，接口失准

1.1 人机交互的基本概念

人机交互，简单定义就是“人机之间的信息交换”。首先让机器理解人的视觉、听觉、触觉和行为，如果机器能够通过这些渠道理解人的活动，那么它就可以为人类提供更为主动的服务。从狭义上理解，人机交互是操作系统的一个组成部分。在下图所示的操作系统中，有负责存储、计算、网络的资源管理部分，负责系统调用的部分；除此之外，操作系统中还有负责交互的部分，就是我们的 UI。其实 UI 才是用户真正接触到的，是人机之间进行信息交换的通道。UI 对计算机在社会当中的应用方式，对用户的使用体验有直接的、极大的影响。

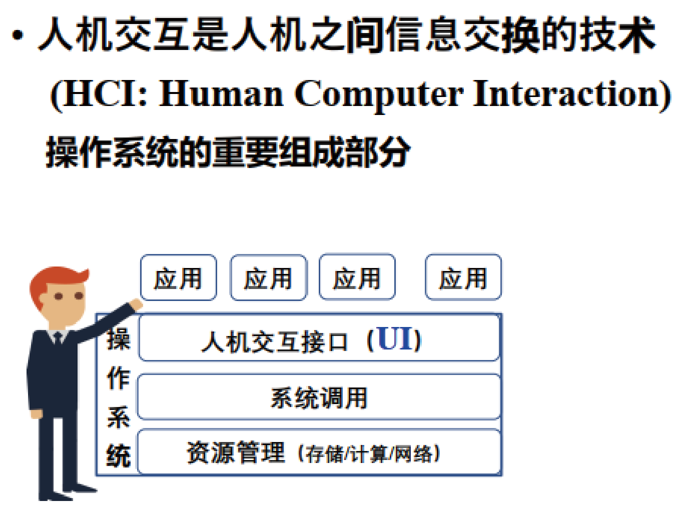


图 1：人机交互是人机之间信息交换的技术

1.2 人机交互发展历史

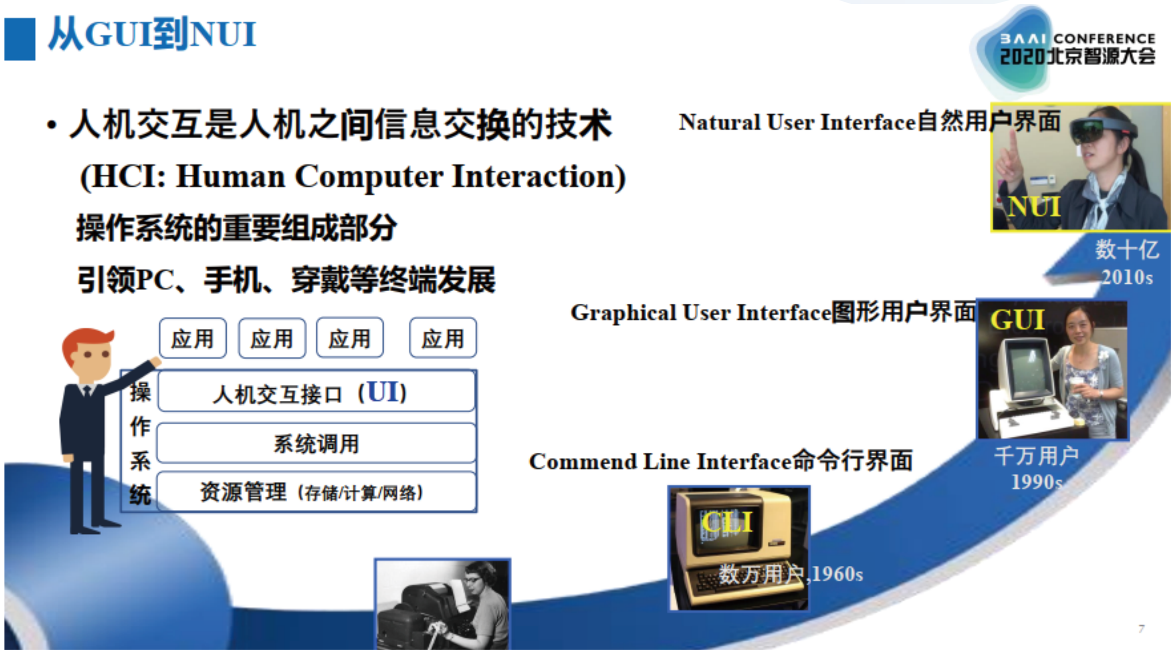


图 2：人机交互发展历史

关于人机交互的发展历史，如上图所示。最初没有所谓人机交互，只能用打孔这类机器的方式和机器交互。六十年代出现了新的人机交互方式——命令行界面 (Command-Line Interface，CLI)，比较接近人的自然语言，实际上还是机器语言。随后，图形用户界面 (Graphic User Interface，GUI) 在八十年代出现、九十年代普及。GUI 的出现真正引发了巨大的革命。从用户的角度，人对图形更容易理解，GUI 这种合适的、简便的交互手段，直接促成 PC 的出现。从商业的角度来看，UI 的发展也推动了互联网的发展。以 GUI 为例，大家比较熟悉的代表人物是 Steve Jobs 和 Bill Gates，很多书本、电影和纪录片都讲述了 GUI 的诞生，例如 Mac 和 Windows，它们都是操作系统中典型的关于 UI 的管理部分，并且具有特定的图形用户界面。

以上是普通人看到的商业上的技术和成果。实际上 GUI 产生的背后是有计算原理以及产业技术的革命作为支撑。史元春介绍了与 GUI 产生相关的 4 个图灵奖得主：Allan Newell、Alan Kay、Douglas Engelbart 和 Ivan Sutherland。这其中，Allan Newell 建立了关于认知和行为之间的行为之间的模型；Ivan Sutherland 在六十年代初提出了关于图形用户界面的图形学；Douglas Engelbart 发明了鼠标； 而 Alan Kay，这次也来到了智源大会，PC 上很多交互模式的提出和技术的实现也秉持着他所提出的 OO(Object Oriented) 的理念。所以，虽然 GUI 今天稀松平常，但作为一个革命性的原理和技术，历史上有很多伟大的科学家参与其中。

到了新世纪出现了自然用户界面 (Natural user interface, NUI)，最直接的例子是 2007 年的触屏手机，然后是 VR 眼镜。NUI 让我们可以不要传统的输入设备，用身体本身就能完成与机器的交互。

NUI 的出现，使得人机交互接口由之前的不好用、不好记变得相对好用，但同时也带来了接口不准确的问题。

1.3 交互方式的革新

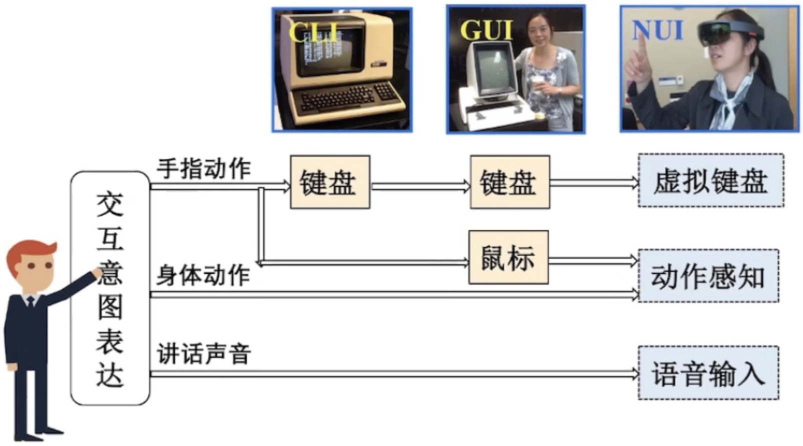


图 3：交互意图表达

再次回顾一下 UI 的发展历史，从命令行界面到 GUI 再到 NUI，人与机器之间交互意图的表达一直在发生改变。命令行界面通过人手指的表达传递给键盘，键盘的命令给了计算机，计算机再解读、反馈给用户。到了 GUI，键盘还是在的，与命令行界面没有太大区别，但是多了鼠标。虽然看着简单，但鼠标背后的原理和技术的贡献是十分伟大的。因为通过一个简单的操作，可以把很深奥的字符命令表达变成图形命令，一个点击的动作就可以操作所有的命令和内容。以上都可以认为是人的手指作为人机交互的接口。

到了 NUI 阶段，鼠标这类实体的外界输入接口都不存在了，触屏手机、Kinect、HoloLens 等设备的出现带来了新的人机交互接口，主要的有三个：虚拟键盘、动作感知和语音输入。

为了输入文本，键盘还是有必要的，但是变虚拟了，比如在手机上变成了软键盘，在眼镜中变成了浮空的键盘。虽然已经不是硬键盘，但是键盘的功能还在。手的点击作为空中虚拟对象仍然在继续。更多的接口变化来自我们自身，我们的动作姿态 (Gesture) 会变成直接命令，跟系统进行交互，伴随而来的是各种各样动作感知的接口。除此之外，NUI 不需要在固定的桌面环境工作，得益于 AI 的技术，语音在移动环境下成为了输入的接口和工具。

1.4　接口不准确

$$B_e(t) = \mathrm{W(UI,\ user\ state)\ x\ log\ (1 + S(t)/N(t))}$$

(专注度、视觉注意、心智匹配、熟练度...)

信道传输率

图 4：信道传输率公式

人机交互作为交叉学科，有一个基础理论来自于信息论，上式为用以表征接口的带宽。带宽受很多因素影响，其中 S(T)/N(T) 作为信噪比，是接口本身的特性。对于 GUI 及之前的接口，信噪比 S(T)/N(T) 都是一个确定性的输入，只有个体在使用状态（User State）上的差异会影响传输效率。

但自然人机交互需要支持新的、更大量的个人应用场景和终端，接口几乎没有准确的硬件，都是身体本身和讲话声音的发音。这样的场景下接口不确定性和信号中的噪声很大，有效信息利用很有限。在这样的前提下，操

作系统中如何支持基本的交互功能显得尤为重要。

图 5：硬件接口不准确的具体体现

接口不准确的具体体现在哪里？有三个方面的问题：1）对虚拟键盘而言，虽然还是需要手指动作操作，但是键盘的设计不符合人体工学；2）对动作输入而言，人的动作含义很多，并且是有意和无意耦合的，需要在连续活动当中把有意的交互识别出来，不需要所谓的唤醒词、唤醒动作。Kinect 这类设备需要有意地设置一个开关，比如挥手，交互才能开始。但是在很多场景下我们是不能要这个开关的，所以需要接口支持动作的判断；3）在语音方面，我们今天利用 AI 技术建立物理模型进行内容识别，已经做得很好了，但现实生活中人在各种场景下讲话的内容，模型对话意的理解还不够，比如重音、情感、与姿态的关系、情境等等。

以上种种问题都反映在接口产生的带宽上，这些需要改变和不准确的地方，怎么才能准确？这就需要对人的自然行为能力进行建模了。

二、建模人的自然行为能力

人机交互最根本的是直接与人交互，如果人的行为能力不能建模、不能计算、不能优化，就不可能通过学习和训练，得到一个新的接口。

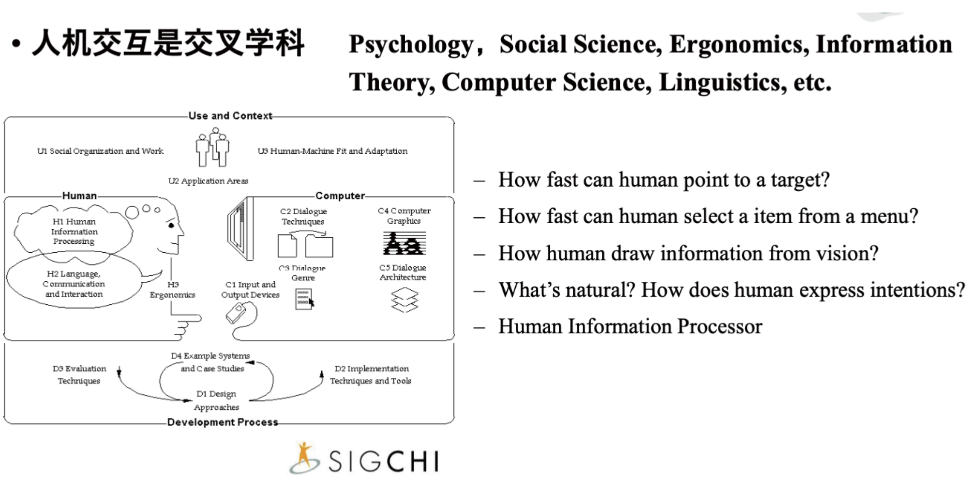


图 6：人机交互是交叉学科

上图是 ACM SIGCHI 给人机交互的定义，可以看到人机交互是一个交叉学科，心理学、精神科学、语言学、信

息理论、人体工学、社会科学等都对人机交互有贡献。上图中列出的一些研究问题，在 GUI 时代已经有比较多的研究，也给 NUI 提供了很好的借鉴和方法。

Allen Newell 曾参与建立的 Human Information Processor Model（HIP Model）有巨大影响。该模型把人的感知行为和认知的子系统建立成类似冯诺依曼结构的模型，如下图所示，每一个子系统的处理器、存储器的参数是通过大量的生理和心理实验获得的，同时该模型还总结了一些非常重要的运行原理。

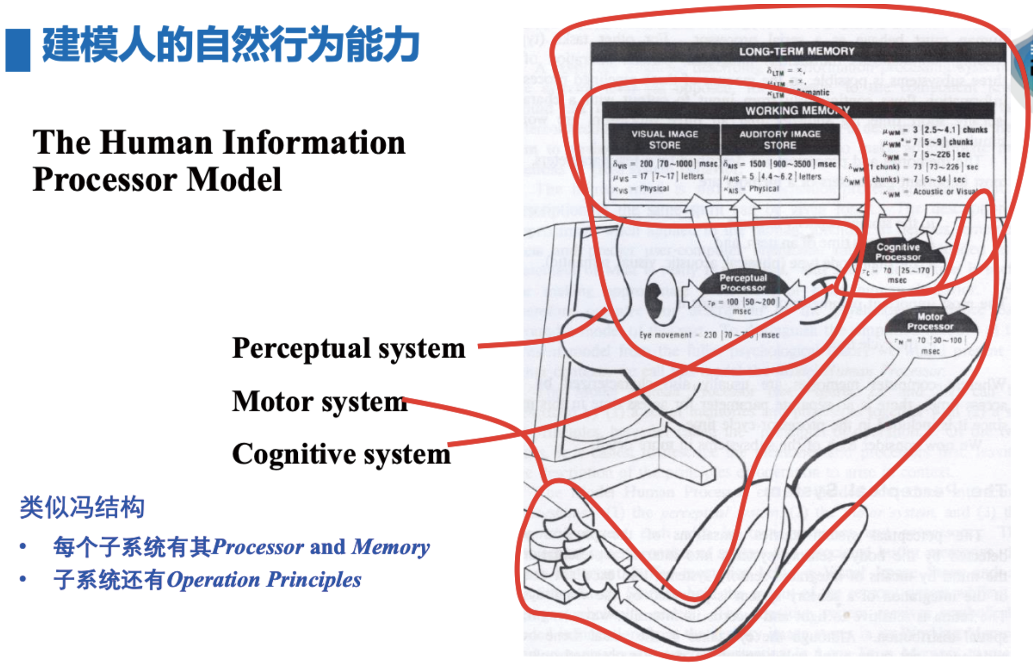


图 7：建模人的自然行为能力

具体以人手点选目标的 HIP 模型为例。在 HIP 模型中，人脑处理手部行为的模块是很大的，这和手的实际体积不成正比，因为手有很强的表达能力。HIP Model 当中对人手点击的速度、精度以及和认知位置的关系做了很深入的研究，公式如下所示。

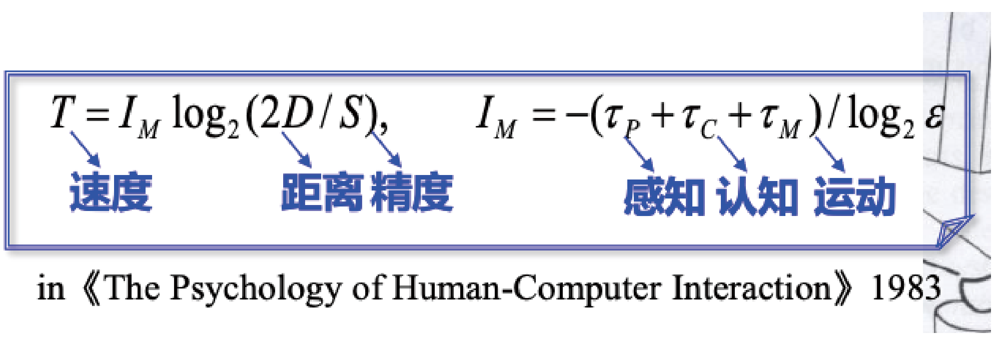


图 8：速度、距离、精度、感知、认知和运动之间的关系

该公式关于速度、距离、精度、感知、认知和运动之间的关系描述得非常精准，并且成为 GUI 的一个优化的基础，在一些真正有生产力和竞争力的接口上都很好地体现了这样的关系。这种建模能力和方法对 NUI 也有很好的借鉴。

三、虚拟键盘胖手指如何快准输入

虚拟键盘，即为软键盘和浮空键盘，其尺寸不符合人体工学设计，也没有触觉反馈。由此带来的结果是输入速度慢、点不准等问题。点不准在人机交互领域叫做胖手指（Fat Finger）问题。在胖手指问题中，放松的状态让人点击触屏上特定位置，正确率可能不到 50%。

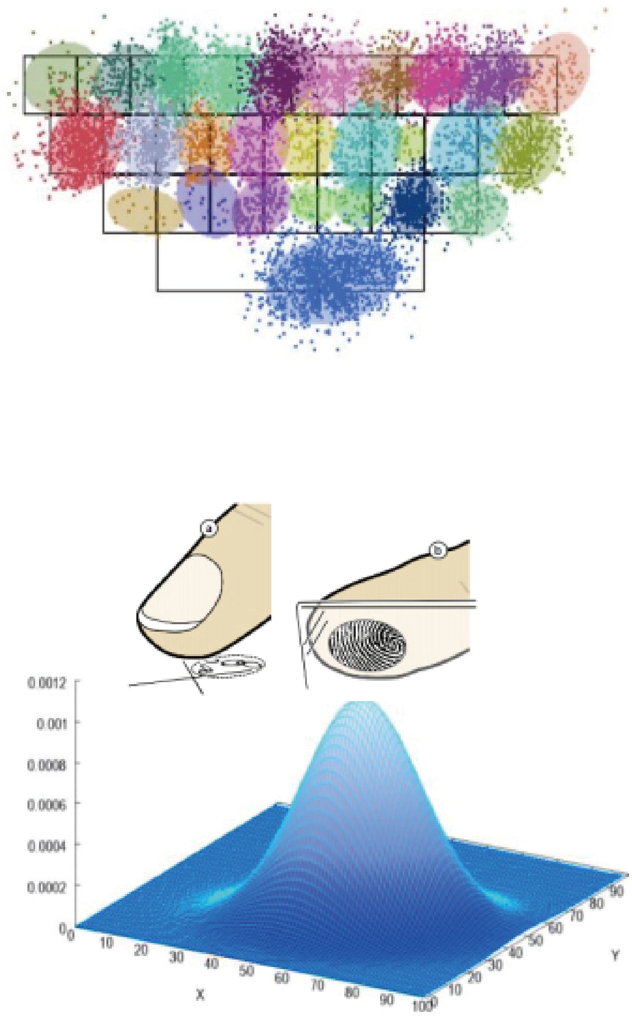


图 9：人机交互领域的胖手指（Fat Finger）问题

史元春团队参考 HIP Model，建立了速度精度双极的手指运动控制能力模型，如下图所示。

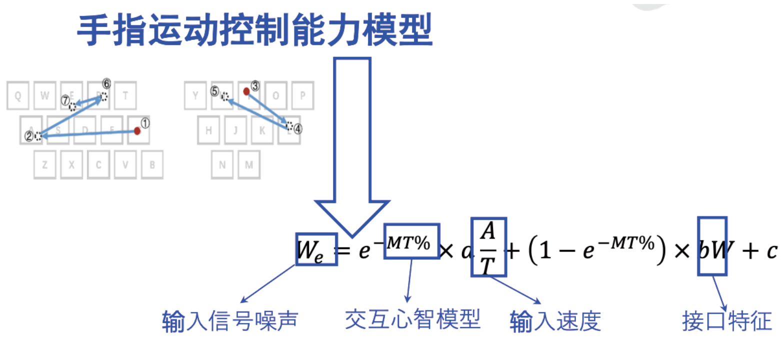


图 10：手指运动控制能力模型

这个模型考虑到了人在交互过程中的使用习惯，也即 Mental Model，因而比 HIP Model 更加复杂。该模型量化了放松输入状态下，输入落点噪声与接口尺寸、输入速度、视觉注意之间的关系，从而实现了虚拟键盘上，手指运动控制能力的可计算、可优化输入纠错能力。

此外，键盘本身其实也是文本输入的贝叶斯解码的过程，如下所示。

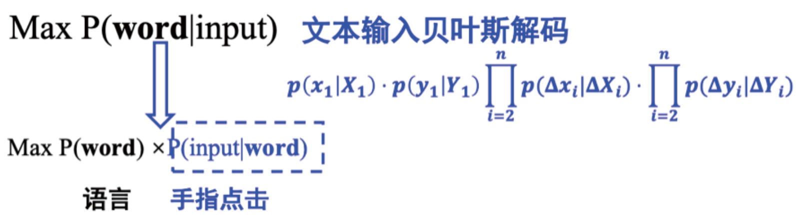


图 11：文本输入的贝叶斯解码过程

上式由两个模型构成，一个是以语料库为代表的语言模型，一个是手指运动控制模型。将二者结合，可以很好地优化点击的准确率，从而在软键盘上提供比较快和比较准的输入法。

史元春团队所提模型的好处在于:

(1) 模型可解释，真正在人的手指控制能力的基础上计算和推理出来;

(2) 利用了先验知识，使用小样本的训练数据就可以支撑该模型;

(3) 模型可以扩展到多种交互接口。

为了验证模型可行性，他们使用一个智能手表上的软键盘作为交互接口，如下图所示，按键只有几个毫米，完全不可能按准，但是经过所提模型优化后的输入速度和准度，基本可以达到大屏手机上的输入速度。该输入法模型实际应用在华为的智慧输入法和搜狗的智慧输入法中，极大地提高了输入速度和准确度。

所提模型的另一个应用是在 AR/VR 眼镜中。这是非常难用的一个场景，普通人每分钟可以在硬键盘上打字 60 个左右，在手机上大概 30–40 个，在 AR 眼镜上实测只有个位数。史元春团队在原有模型基础上建立 BlindType 输入模型，靠每个人已有的肌肉记忆和远端视频上的反馈进行交互输入。在实验中，人不需要看手

机，只是需要大致位置，和语音模型，最后可以实现每分钟 20 个字的输入。

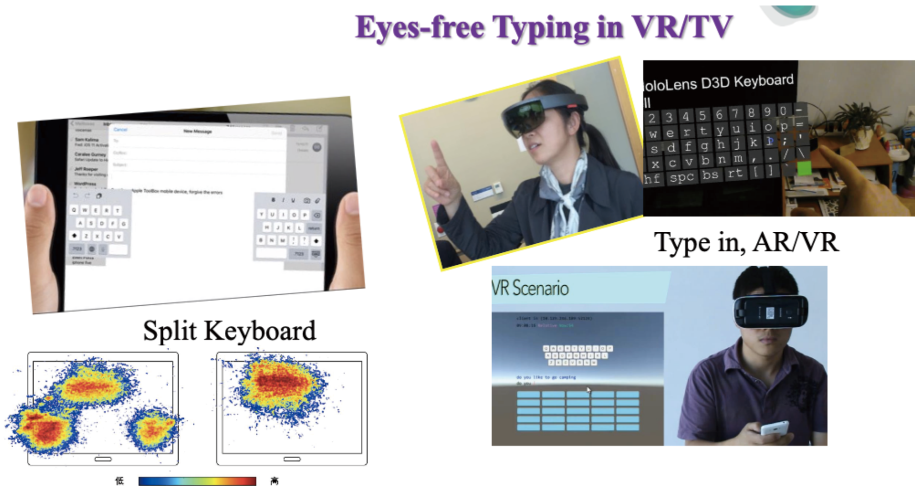


图 12：Eyes-free Typing 应用

同样的 Eyes-free Typing 应用在 iPad 的 Split Keyboard 上，在原有输入速度基础上提高了一倍多，验证了方法的有效性。

史元春团队还为盲人设计了一款键盘。下图图中左侧是明眼人键盘，可以实现单词级别的纠错，右侧为盲人键盘，配备有读屏的软件，但只能做初级的输入。史元春团队通过实验，估计了手的相对位移，并对手指和字符位置之间的关系做了大量的统计，建立精确的模型。优化后的键盘会产生微小的变动，虽然输入的位置错了，但是模型预计到了想输入的内容，可以把键盘移动到他的手指下面，使用者听到的就是输入正确的结果。

图 13：世界首款盲人键盘

这也是 5 月 21 日清华和搜狗发布的世界上第一款盲人键盘，纠错率提高了 60% 以上。还和中国残联、盲协做了测试，原理性的论文获得了去年 CHI 会议的最佳论文提名。

四、连续动作如何准确识别动作意图

史元春认为，连续动作的输入更具挑战，因为动作有意和无意是连续的，模型需要准确识别用户的动作意图，避免误触发。难点在于动作之间的连续性、随机性和动作传感的不完整。

针对此问题，他们提出了运动参数时序模型，如下图所示。

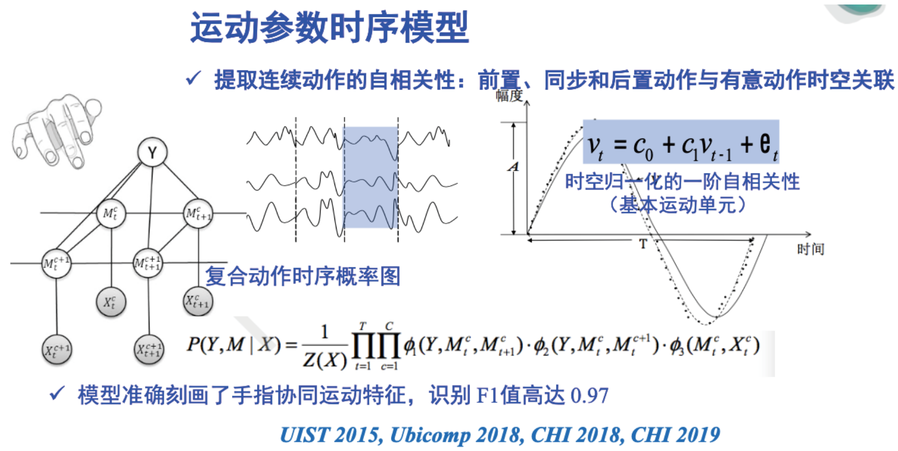


图 14：运动参数时序模型

该模型提取连续动作的自相关性，即有意动作与其前置和后置动作之间的时空关联。因为人是有结构的、有基本运动单元的基本参数，在归一化上可以建立自相关模型。同时复合动作是靠多个子动作的时序概率图构建起来的。以手的协同运动特征为例，该模型在特定任务上的识别 F1 值可以达到 0.97。动参数时序模型在具体接口上具有良好的表现。

AR/VR 眼镜需要依赖视觉辅助进行动作选取，即只有在眼镜的可视范围中看到了物体，才能进行下一步的选取动作。这种与空间对象的交互方式存在一定的问题，一方面视觉注意需要花费一定的时间，另一方面这也会增加疲劳度，甚至造成使用中的眩晕感。

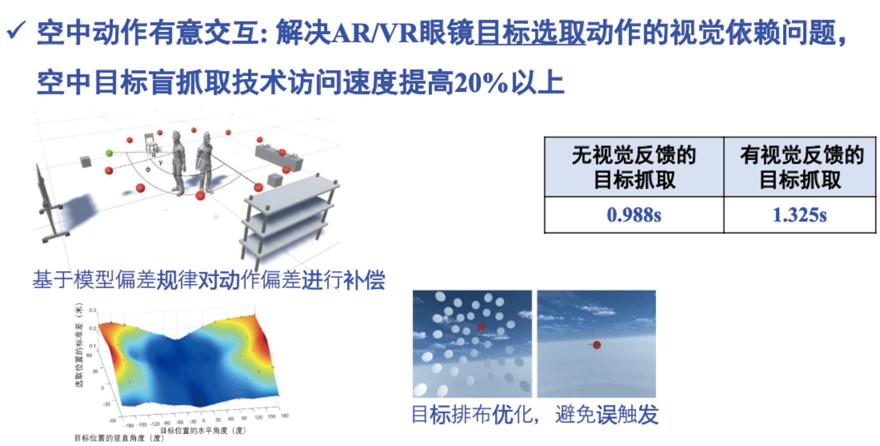


图 15：解决 AR/VR 眼镜目标选取动作的视觉依赖问题

针对上述问题，史元春团队实现了空间的盲操作，即针对空中目标，在建立运动参数时序模型的基础上，基于模型偏差规律对动作偏差进行补偿，从而实现不需要用户注意力的盲抓取。在进行的大量抓取目标实验中，抓取速度明显提高，准确率与视觉依赖条件下基本相同，都是 98% 左右。该方法已经广泛应用在 AR/VR 眼镜中，效果十分出色。

为进一步解放双手，史教授团队还借助头的运动设计了一系列的基本操作，如下图所示。

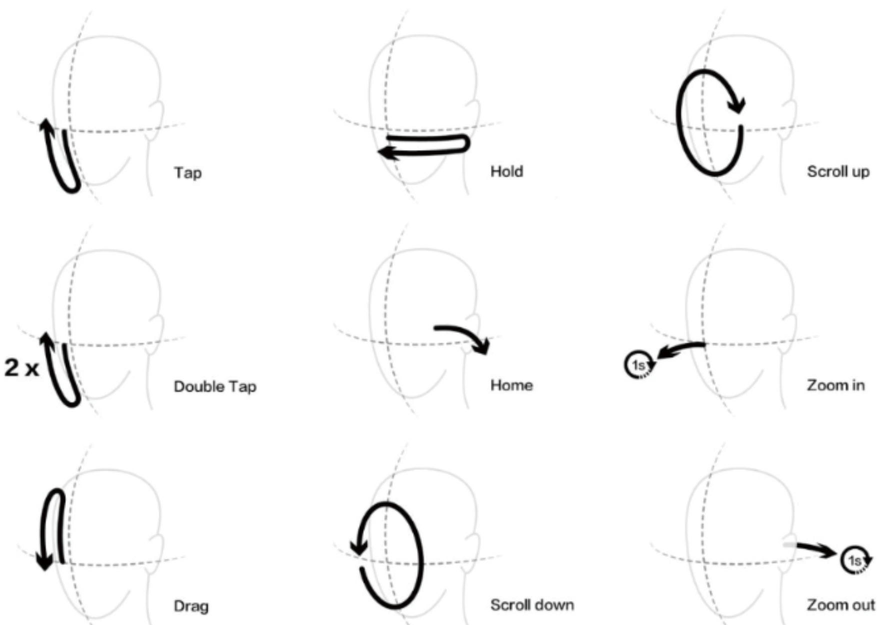
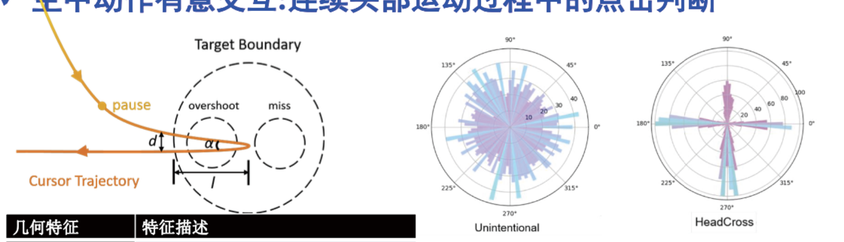


图 16：借助头的运动设计的一系列基本操作

下图给出了连续头部运动过程中用于判断的时序特征:

✓ 空中动作有意交互:连续头部运动过程中的点击判断



几何特征	特征描述
进出距离	光标进入和离开目标的位置之间的距离
始末距离	动作开始和结束时光标位置之间的距离
最大转角	光标在目标内最大转角的大小
最大曲率	光标轨迹中最大曲率的数值
线性程度	以最大转角为界两侧轨迹的线性程度
目标内长度	光标在目标内轨迹长度
转角一致性	最大转角处与目标内最远距离处的差别

时序特征	特征描述
目标内时间	光标在目标内时间长度
速度低谷序号	光标在进入目标前，在目标内和离开目标后的速度局部最小值的帧序号
速度峰值序号	光标在进入目标前，在目标内和离开目标后的速度局部最大值的帧序号

图 17：连续头部运动过程中的点击判断

在手机上面也有意图判断问题，最突出的是曲面屏的误触问题。在曲面屏手机出现早期，无法在软件上区分有意和无意触碰动作，极大地影响了用户的使用体验。史元春团队和华为合作，基于运动参数时序模型实现了高精度的握姿识别算法，降低了误触事件的发生概率。

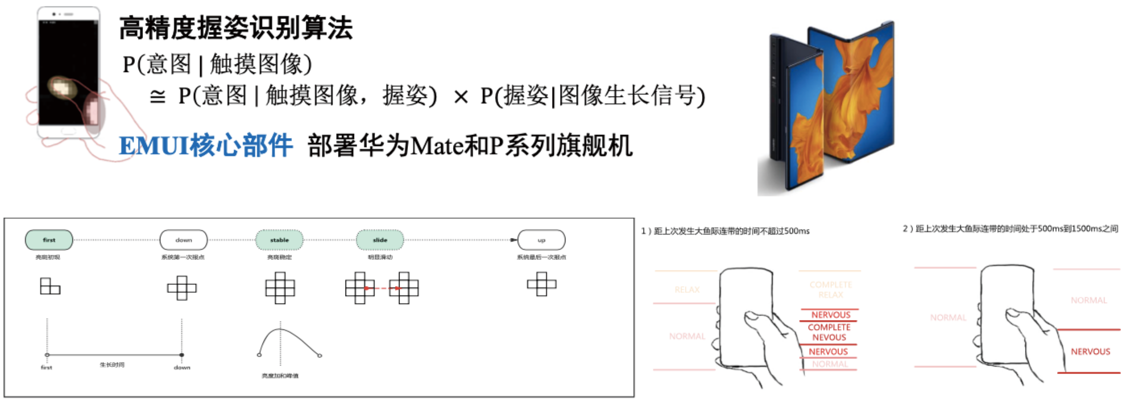


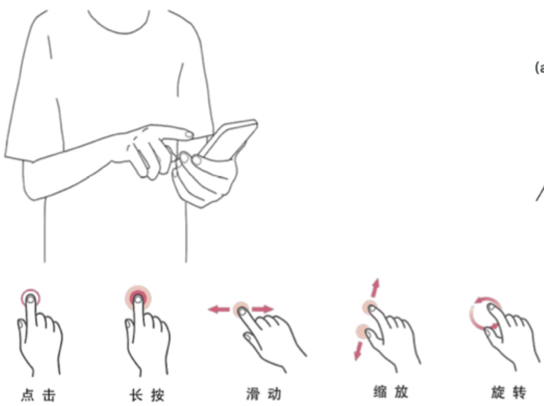
图 18：高精度握姿识别算法

此外，史元春团队还基于该模型设计了针对盲人用于的耳势交互模型。

图 19：耳势交互模型

最后，史元春教授还介绍了他们团队在全手型交互方面的进展。全手型交互，意味着手机可以看到用户用于交互的手，空间中手指做出的即时动作都可以被直接识别。如此一来，可以将交互空间从二维平面扩展到三维空间上。史元春展示了一些目前的实验结果，如通过手势选菜单、通过手势玩游戏、通过手势拍照等。

✓ 手机全手型交互



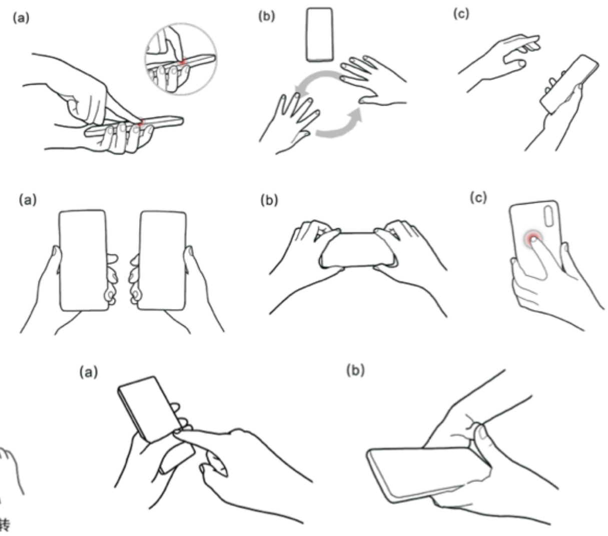


图 20：手机全手型交互

综上所述，人机交互考虑的是编码，编码过程需要符合以下要求： 利用人的学习和记忆的资源越少越好，但是解码的过程需要研究者从计算机的角度还原人的意图，获取准确的动作信息。

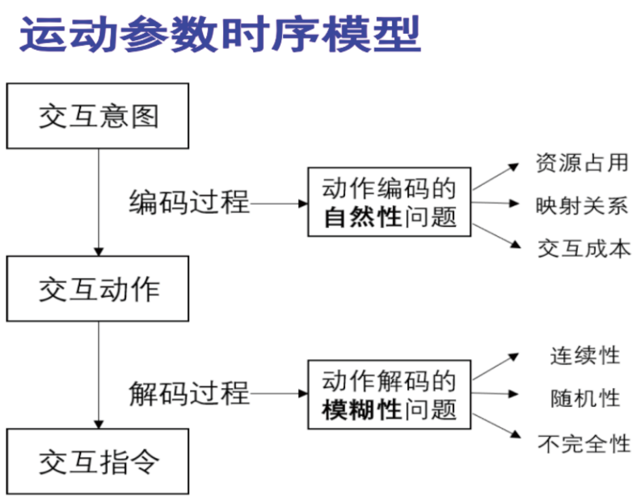


图 21：运动参数时序模型

智源研究院院长黄铁军：脉冲视觉：重塑视觉信息处理技术体系

整理：智源社区 蒋宝尚

在第二届北京智源大会“机器感知论坛”上，智源研究院院长、北京大学教授黄铁军做了《脉冲视觉：重塑视觉信息处理技术体系》的主题演讲。

在本次演讲中，黄铁军从两个角度介绍了当前计算视觉体系的误区。其中，在“数码相机的 Bug”部分提到，当前的相机“技术”简单地、直接地继承了胶片时代的视觉信息处理的模式（图像和视频），非常不适合当前的技术发展趋势；在“计算机视觉的误区”部分提到，传统的“Video 摄像头 + 计算机 + 算法 = 计算机视觉”的定义是错误的，无论是 Video 摄像头层面还是算法层面，都需要进行变革，摄像头需要突破传统数码相机的 Bug，而算法需要借鉴生物神经元的性能。

以下是智源社区编辑对黄铁军演讲做的文字整理：

今天的报告题目是《脉冲视觉：重塑视觉信息处理技术体系》，主题内容是对当前计算机的基本概念、基本做法进行重新思考。

整个报告的提纲有两个部分，第一部分是数码相机的 Bug，第二部分是计算机视觉的误区。

一、数码相机的 Bug

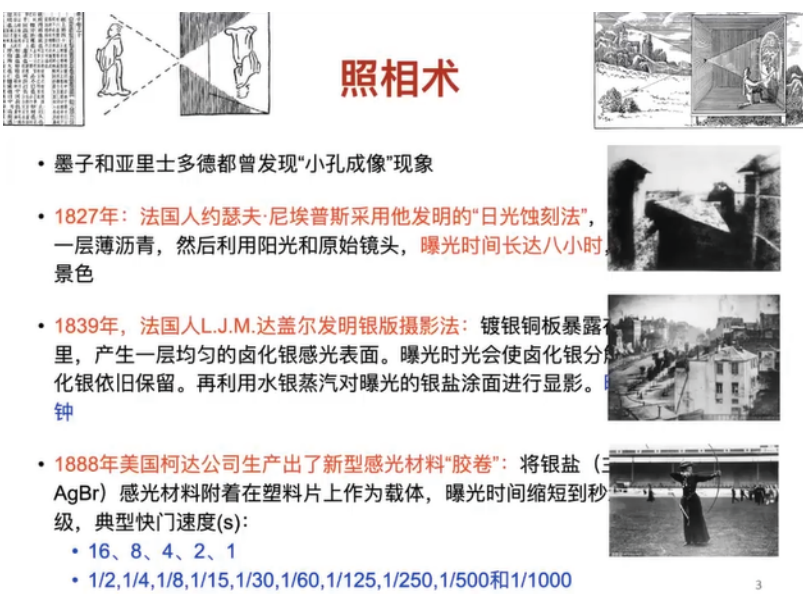


图 1：照相机发展史

现在开始第一部分，数码相机的 Bug。这里的 Bug 概念和计算机视觉研究员为了完善系统而进行寻找的 Bug 同义。

那么，为什么数码相机有 Bug？其实，这主要来源于人类的惯性思维。数码相机的前身就是模拟相机，实际上今天数码相机的各种各样的观念是来源于模拟相机。

下面从“人类的第一张照片”开始解释这一“来源”。第一张照片拍于 1827 年，拍摄者是法国人约瑟夫·尼埃普斯，采用的技术是他发明的“日光蚀刻法”，具体操作是用阳光照射一层薄沥青，通过把沥青晒软就能得到薄厚不同的“影像”，其中曝光时间长达八小时。因此，可以看出当时的技术非常原始。

到了 1839 年，在化学技术进行和工业革命逐渐发展的时代背景下，拍照技术也不断发展，最直观的表现是：我们拍照每一幅图像的曝光时间不断减少，从最开始的“秒级”到 1/2 秒再到 1/4 秒等等，现在甚至达到了 1/1000 秒。其实，上述提到的概念，随着模拟相机的使用便被固化下来，因此今天的数字相机也在用同类的概念。

图 2：人类第一个程序员

举一个更为生动的例子，如上图所示，这位女士是人类的第一个程序员，ADA 编程语言的命名就来自于她。在 1845 年时，为这位“女士”拍下这幅照片大概要花费要半个小时，换句话说，她需要安安静静地坐上半个小时。

所以，任何一幅照片的形成，都是以时间为代价的，只不过在技术不成熟的早期，其时间比较长，当前相机需要的曝光时间比较短，但这也不意味着不需要时间。

图 3：Video 视频发展史

另一方面，视频也并没有比图像更加“高明”，因为视频是图像的“序列”，例如一秒钟视频包含 24 帧图片。

其实，从当年胶片电影“变成”电视之前，已经犯了一次“错误”，例如当时用的词是 Video，这意味着所有的“东西”都是在电信号的层面进行处理，但实际上 Video 表示视觉信息的方式还是电影胶片，也即每秒钟多少帧图像。

这背后带来的思考是：当我们不再依赖于化学胶片的时候，是不是还要用每秒钟几十帧的图像表示视觉信息?

而且，在一开始，这种视觉信息表达方式就有一个根本性的矛盾，矛盾点在于曝光成像中曝光时间Δ t 的如何选择。

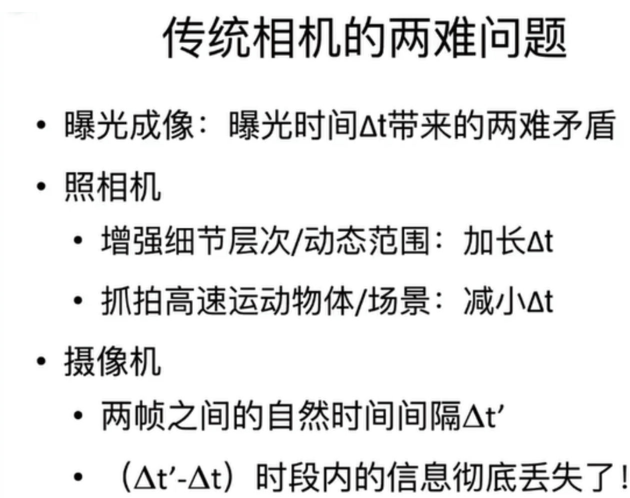


图 4：传统相机的两难问题

例如，在照相机中，如果想增强细节层次或者动态范围，所采取的是加长Δ t 策略； 如果想抓拍高速运动物体或者场景，所采取的是减小Δ t 的策略。因此，曝光时间带来的矛盾正影响着我们视觉信息的获得。

另外还需要明确的是： 在拍摄视频的时候，提到的每秒钟包含 30 帧图片，并不意味着 2 帧之间的 33 毫秒的“时间”会成为一幅图像，因为 33 毫秒的时间太长了，会导致图像模糊。真正拍摄时通常只会用几毫秒（例如 3 毫秒），剩余的几十毫秒（30 毫秒）都“浪费”了。

换句话说，在拍摄过程中，只有 1/10 的“时间”被记录了下来，剩下 90% 的“时间”并没有记录，所以摄像机作为一种记录视觉信息的方式，每秒钟 30 帧的图像不是完备的，有 90% 的信息被迫丢失。

通过上述的分析，然后回过头来思考当前视觉信息处理的方式，我们图像视频的概念可能是错了，因为这是胶片时代遗留的做法，现在是电子时代和数码时代，我们还那么愚蠢地每秒钟拍几十帧图像，这种做法本身是错误的。

那么应该怎么做呢？正确的做法是把光完整地记录下来，也即一个理想的相机应该把每一个光子到达的时间和空间以及每个光子的频率记录下来。这样才能达到物理上完备的信息采集。

虽然普通的 Sensor 电路处理无法跟上“光的脚步”，但是也不能退化到传统胶片模式。因此我造了一个词叫做视达（Vidar=Visual Radar），目的是为了和 Video 进行区分。

视达（Vidar）的视觉信息当中每个单元叫做 Vit 视元，如果相机能够达到极致，一个 Vit 就是一个光子的到达。当然，现实当中的相机肯定做不到这么精密，那么可以把一群光子到达的时间变成一组，用一个比特进行记录。其中，每个感光器件视为一个光子捕捉器，当收集的光子能量达到约定阈值时，就产生一个脉冲，这个脉冲及形成这个脉冲所持续的时长称为一个视元。

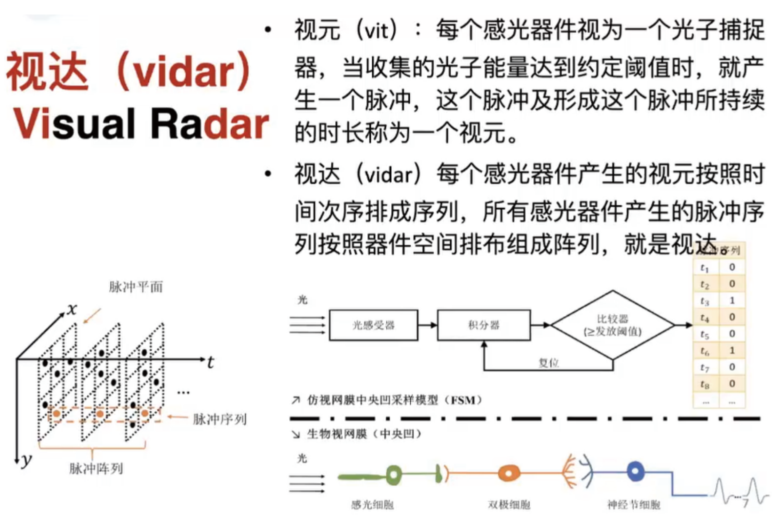


图 5：视元和视达的定义

因此视达（Vidar）的准确定义是某种阵列，即每个感光器件产生的视元按照时间次序排成序列，所有感光器件产生的脉冲序列按照器件空间排布组成阵列。

基于视达概念得出的相机，能通过微秒级的时间精度进行采样，从而得到的一个高速的比特流。另外，这个比特流要比原来的视频更接近物理的现实，因为这个比特的产生代表着前面一段时间的光子积累达到了一个阈值。

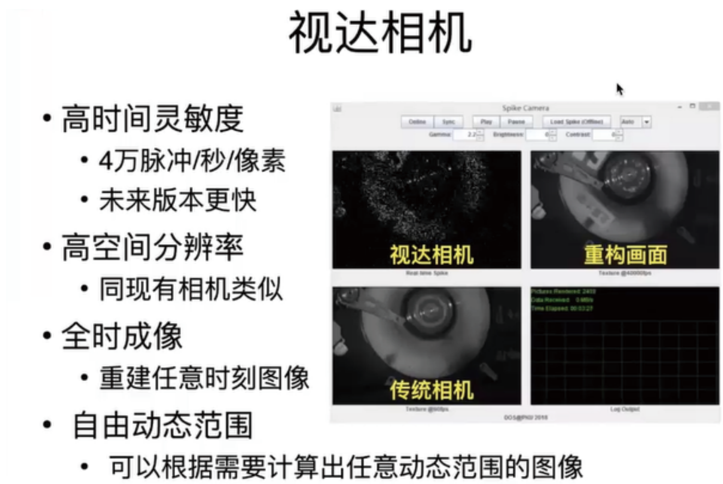


图 6：视达相机的特点

这样做出的相机有什么用呢？举例而言，我们第一款相机每秒可以产生 4 万个脉冲，可以抓拍快速运动的物体。例如在拍摄硬盘旋转（转速高达每秒钟 7200 转）的过程中，硬盘上面的文字都能够清晰可见。

另外，为什么相机拍摄的是比特序列，呈现的确是灰度图像呢？原理很简单：可以把比特想象成光子，而从一组光子里面形成一组图像简直太轻松了，其中可以采用的算法是：考虑两个比特之间的时间差，如果时间差越长，那么比特很密集，也就说明光很亮。如此便能把亮和暗的灰度区分开来。

而更巧妙的地方在于：我们可以根据视达（Vidar）比特流产生任何图像。其实这种“巧妙”对于视觉信息记录非常关键，因为我可以得到连续的信息表示。

因此，为了更加明确上述概念，我定义了两个词：全时成像（Fulltime Imaging）和自由动态范围成像（Free Dynamic Range，FDR）。其中，每个视元的持续时间和光强成反比，据此可以得到该像素位置任意时刻的光强，从视达中得出任意时刻图像，实现全时成像；如果时间窗口扩大到多个视元，则可以得到更高动态范围的图像，称之为自由动态范围成像。

从上述定义描述的“场景”可以看到，一旦相机变成这种形式，其拍摄效果比传统每秒几十、几百帧图像的“摄像”要好的多。这也意味着任何人都可以拍出来高素质照片，或者对日常人眼都看不清楚的东西进行抓拍。

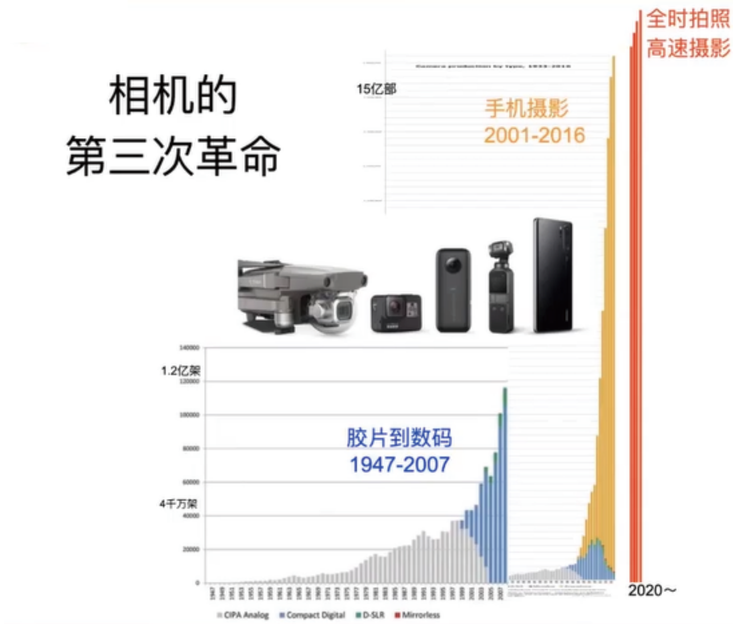


图 7：相机的发展历史

同时，我也认为视达会导致相机的第三次革命。如上图所示，相机的第一次革命发生在 2000 年左右，当时相机数码化了，相机的数量从每年年产量 4000 万台左右提高了 3 倍，差不多达到了 1 亿多台。第二次革命发生在手机时代，相机数量变成 15 亿部，从数量上来看，相对于数码时代，手机带来的革命是 10 倍的增长；从今年开始，将会迎来相机的第三次革命，因为全时拍照，高速摄影意味着能够重构任意时刻的图像。

总结一下，数码相机的 Bug 具体指的是：当前的相机“技术”简单地、直接地继承了胶片时代的视觉信息处理的模式（图像和视频）。因此，数码相机设计应该按照数码的方式进行，让光流变成比特流，从而成为将来的趋势。

二、计算机视觉的误区

计算机视觉的学科方向是上个世纪八十年代建立起来的，其中，基本概念在六十年代就提出来了，当时 MIT 人工智能实验室也在筹建，筹建的前期有一个项目叫做计算机视觉，这个项目也被追认为计算机视觉的第一个项目。

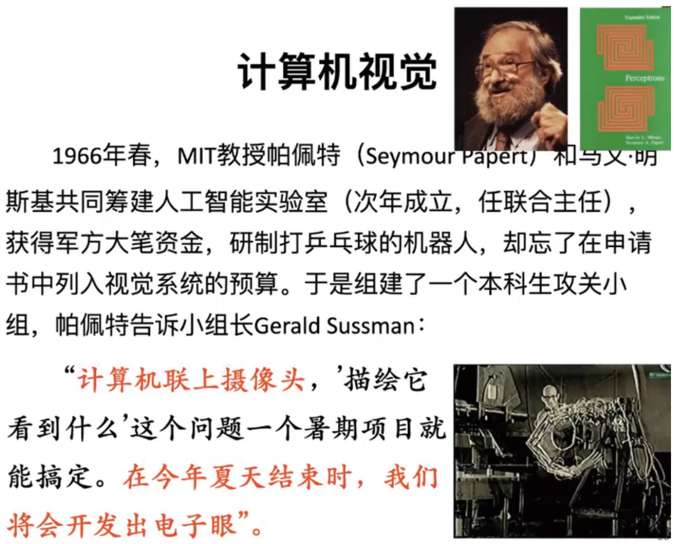


图 8：计算机视觉

当时，也就是在 1966 年，对计算机视觉的想法和实践都围绕一个问题，即：计算机连接摄像头能够看到什么？为了解决这个问题，人们下意识的可能会需要一个摄像头，需要一台计算机，然后写程序写算法，识别处理摄像机拍下来的每秒钟几十幅图像，然最后进行检测跟踪识别等等。

但是，这种做法无疑对计算机视觉、机器视觉做了范式上的定义，即将其等同于 Video 摄像头 + 计算机 + 算法。然而这样一套技术路线是不是正确的？这样一套做法从头到尾都是错的，在前面一部分的讲述中，就已经明确知道“摄像头”是错误的（因为它根本没有把光的信息全部采下来），这背后代表的是输入错误。

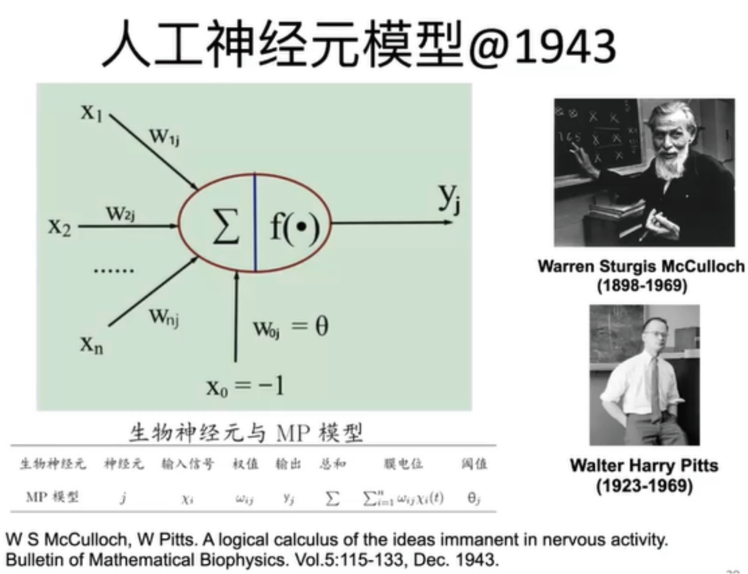


图 9：人工神经元模型

另一方面是算法层面的错误，其根源不是某个人的问题，而是人类这一个群体犯的错误。当然，这个错误并不是一开始就犯的，早期研究者认真地在研究生物视觉，比如今天用的深度网络当中的神经元模型。当时 Warren Sturgls McCulloch 和 Walter Harry Pitts 两位就研究过视觉系统，另外也研究过青蛙的视觉。

其实，回顾五六十年代，甚至七十年代，研究者在生物视觉方面做了很多很精品、基础性的工作。代表是 1971 年，德国马普学会生物控制论研究所创始所长赖夏特（W. Reichard）和托马索・波焦（T. Poggio）发现了家蝇视觉飞行控制系统的秘密，并给出了精确的定量描述。

马尔视觉计算理论

大卫·马尔 (David Courtnay Marr, 1945-1980) 1966 年从剑桥大学获得硕士学位，本科和硕士专业都是数学，1972 年 获得生理学博士学位。1973 年，加入明斯基和佩帕特的MIT人工智能实验室。

“我也曾相信，真理从根本上是属于神经的，研究的中心任务就是对神经系统的结构作彻底的功能分析”，**转变为**“用神经元（除作为实现一种计算方法的手段外）对视觉现象所做的任何解释已经不堪回首了。取代它们的便是对一系列问题的明确认识：要计算的是什么东西？怎样才能进行这种计算？计算使用的方法基于哪些物理假设？对可执行这种计算的算法怎样进行分析？”

《视觉：人对视觉信息的表征和处理的计算研究》

图 10：马尔视觉计算理论

到了八十年代，马尔出版了《视觉： 人对视觉信息的表征和处理的计算研究》这部经典著作之后，生物视觉的研究风向全部变成计算机视觉作为主导思想。其实，马尔本身是做生理学博士，但是博士后到了 MIT 一下子转成了计算机，某种意义上是正确的选择。但彻底抛弃生物视觉确实是犯了一个巨大的错误，因为其把学科方向引导错了。现在来看，马尔从神经生理到计算的这个弯转得过急了。

关于马尔的这次转变，特伦斯・谢诺夫斯基在他最近在新书《深度学习革命》中回忆到：“他（马尔）追求一种自下而上的策略从视网膜开始入手（在那里光被转换成电信号），并探求视网膜中的信号如何编码对象的特征，以及视觉皮层如何表示物体的表面和边界……具有讽刺意味的是，尽管马尔在他的视觉研究中采取了自下而上的策略，即从视网膜开始并对视觉处理的每个后续阶段进行建模，他的著作却以倡导自上而下的策略而闻名——首先对要解决的问题进行计算分析，然后构建算法来解决问题，最后通过硬件来实现算法。然而，尽管这可能是在解决问题后对问题进行解释的种有效途径，但对于揭开大脑秘密却算不上是个好方法。

到了 2012 年，深度学习将 ImageNet 视觉对象分类任务的正确率提高十多个百分点，一夜冲垮了计算机视觉积累近半世纪的技术体系，2015 年更是将正确率提高到与人类相当，成了计算机视觉的主导思想。

而在过去的五年，我们一直在追求对生物视觉精细生理过程的仿真，也即想知道光到达视网膜以后是怎样对光进行加工的；想知道大脑是如何对电信号进行处理的。

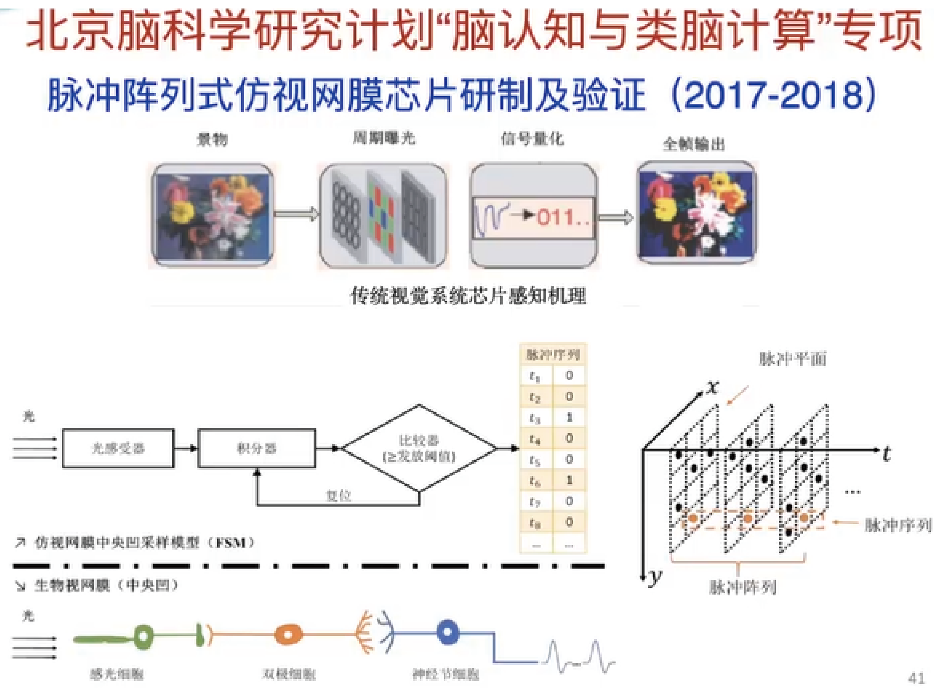


图 11：脉冲阵列式仿视网膜芯片研制及验证

上述研究思路，其实从 2017 年到现在，在大概三年多的时间里已经有了一些进展，例如将生物原理做成芯片。但是对生物原理进行了大幅度进行了简化，原因是生物过程比较复杂，不做简化无法达成共识。

有了芯片和相机之后，能否完成检测识别等视觉任务呢？在过去一年多的时间内进行了实验，值得一提的是，因为我们的相机拍摄的是脉冲流，所以使用的是脉冲神经网络，而不是传统的机器视觉的算法。

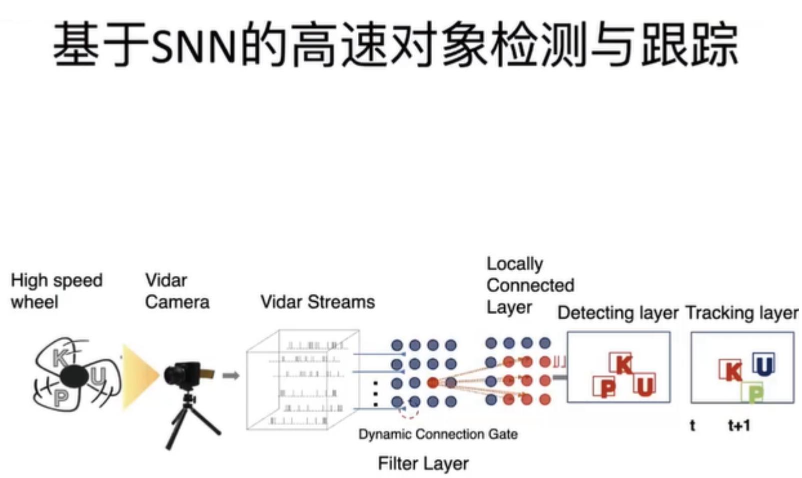


图 12：基于 SNN 的高速对象检测与跟踪

具体模型如上所示，检测跟踪的模型纯粹是人工模型；识别是用类似于深度学习的方法进行，现在也已经实现了实时高速的识别。

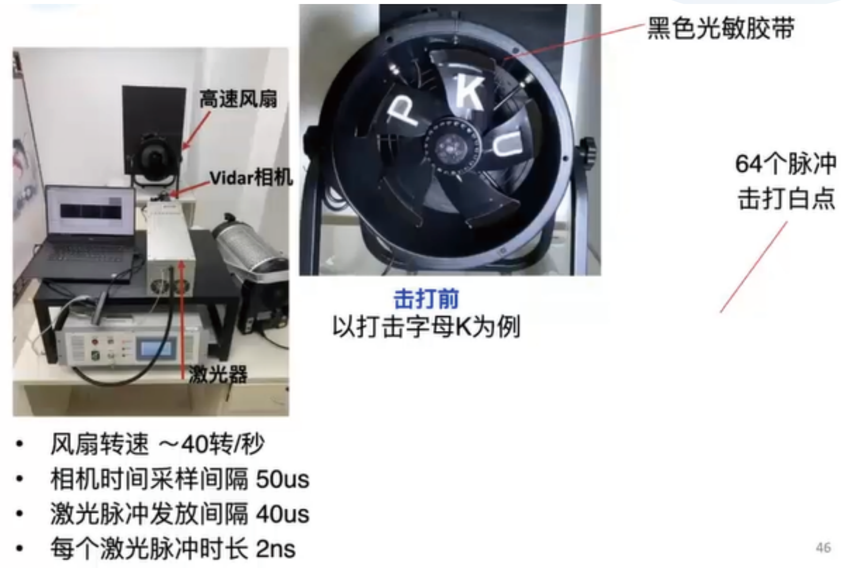


图 13：“相机”识别实验

那么如何证明“相机”不仅能够拍摄到，还能进行识别。我们也进行了一个实验： 在扇叶上贴了三个字母，字母的边上贴了光敏胶带，笔记本上可以进行检测跟踪识别，识别难度相当于“采集”传统每秒钟 4 万帧的信息。结果是：只用一台笔记本运行脉冲网络算法，就可以可以实时识别 PKU 三个字母。

目前只能在笔记本上进行任务，将来要变成脉冲电路，因为变成电路之后就不受主频速度的影响，因为光进入“相机”就变成了比特序列，比特流动的速度就是电流动的速度，前面是光，后面是电，如果打破那些都是人为制造的障碍，速度就不再是一个问题了。

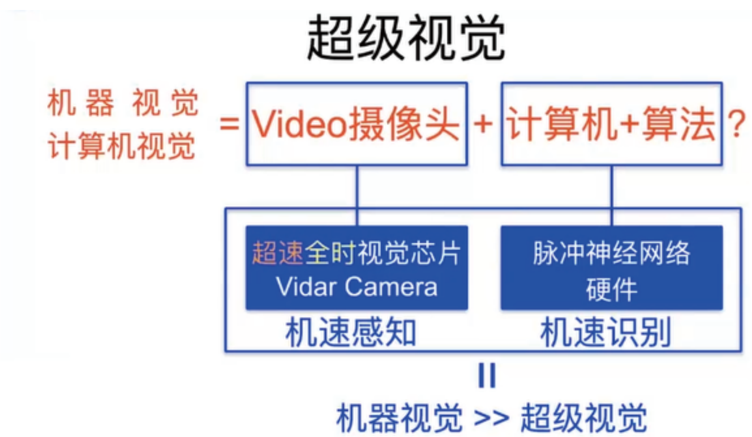


图 14：机器视觉与超级视觉

如此，便能够超越原来基础上图像视频的概念，从而开创一个速度不受限制（只受光速制约）的新技术体系。只有这种打破常规的系统才是未来的机器人真正需要的，才是机器智能的未来。

圆桌论坛：机器感知的新突破口在哪里？

整理：智源社区 贾伟

在第二届北京智源大会“机器感知”专题论坛中，北京大学教授黄铁军、清华大学计算机系教授史元春、北京大学计算机系教授张大庆、清华大学教授孙富春、北京大学教授吴玺宏、中国科学院自动化研究所研究员王亮、中科院计算所研究员山世光共同参与了题为“机器感知的新突破口在哪里？”的圆桌论坛。

以下是要点整理。

黄铁军： 接下来我们讨论“机器感知的突破口在哪里”这个问题。现在做机器视觉的人太多了，很多人号称“机器感知的问题已经解决了，是时候从感知时代进入认知时代了”。

我们认为这种说法，纯粹是对感知的难度了解不足的人说出来的。事实上，感知的问题远远没有解决，并不是机器的识别率有多高，就叫做视觉问题已经解决了。举例来说，现在一个机器人到自然环境里走一走，相比一个低等动物都笨太多了，最主要的就是因为感知系统问题还远远没有解决。所以机器感知可做的东西还很多。

但是，尽管我们都知道机器感知还有巨大的研究空间，以及有许多问题要解决，但是要想说服市场往这个方向走，我们总得有些新的突破口，改变当前深度学习所带来的刷榜潮流，进入一个新的阶段。

这是今天想请大家讨论的。我们成立论坛的时候就有一个基本的理念，凡是用深度学习做视觉的，都不会支持，因为做的人太多了，很多都是没有太大价值的研究。

张大庆： 我很同意你的这个观点。我觉得感知有感知的任务，认知有认知的任务，不能说感知做完了该认知了。至于感知未来的突破口，听了你刚才讲的，我很受启发，感知肯定是像人一样，属于多模态的。今天我们办这个论坛特别好，大家探索不同模态当中感知的极限是什么，这是我们需要真正探讨的。就像刚才你说的一样，传统的相机、摄像头的原理可能不是最好的，后面处理很费劲，识别率有一定的限制。我们做无线信号的，其实也是一样。最近一个重要的目标就是，要研究各种无线信号的感知机理是什么，感知极限在哪里，将来究竟该用什么样的无线信号，什么样的形状、什么样的频率、什么样的带宽等。只有从最基础的角度理解每一种感知模态的极限和理论基础，才能把这个事情做好。

孙富春： 实际上感知应该就是这个过程，我觉得应该从三个方面来讲： 第一就是前瞻，也就是物理载体，我们叫做载荷，怎么获取外界的东西； 第二是后端的处理； 第三是机理。至于如何从生物认知的角度重新认识视觉，实际上是机理层面的重新认识。

山世光： 刚才说感知的问题没有解决，我非常同意。现在大家并不完全是从感知做到认知，而是在做感知的时候把认知的一些东西引进来，但还是在做感知。当然，现在也有一些人确实是去做认知，我认为至少从我的角度还是想做感知，但会把一些认知的东西引过来作为工具来做感知。

为什么大家觉得感知的问题好像已经解决了？我觉得可能是因为大家觉得人脸识别已经解决了，所以感知的问题就解决了，这是错误的根源。之所以这样，可能是大家觉得人脸识别都能够解决，其它的大千世界那么多物体识别的问题，应该都解决了，只需要有足够多的钱，足够多的人，可以去采数据，针对每个不同的物体来做算法，通过大量的机器去做训练，沿着这个技术路线做，好像所有这些问题都可以解决。

事实上，这不是一个可行的技术路线。因为大千世界这么多的东西，不可能有那么多的人、那么多的钱全部靠这种办法去做。所以问题是，现在大家需要重新定义感知的问题到底是什么，我们需要做什么样的感知的问题。

因此，我不是很同意深入学习的感知都不应该做。重要的是，我们的问题到底是什么？要是机器学习能够解决的话，还是应该去做。

我们现在遇到的困难是，我们没有把问题定义清楚，所以导致大家好像觉得好像机器感知已经解决了。

黄铁军：其实人脸识别并没有解决，现在那么多造假就证明人脸识别没有解决，而是能够通过一些手段骗过去。

山世光：这是另外一个问题。所以还是要看问题定义是什么。

黄铁军：之所以能够造假说明神经网络还不够复杂，不能识别那些变化。

山世光：数据足够多的话可以搞得很准。

史元春： 比如隐私保护情况下的人脸识别，很低的分辨率是不是也可以做？因为对人来说很不清楚，一定能够知道那个就是谁，没有更高的分辨率，不需要很大的数据。

山世光：那是非常熟的人才可以，所以还是问题定义。

史元春：所以空间还是很大，但是需要定义目标，不是给定方法。

王亮：“新突破口”这个词我不敢用，但我觉得有两个方向应该提。一个是多模态的融合，一个是多机制的融合。

大家都是做语音、图像或触觉的，这些都属于“感”的方面。现在大家做得都比较独立一点，但每个模态包含的信息源是不一样的，就像人在感知外部世界的时候，有的时候是单一模态，但大多时候是多个模态。模态之间存在互补性，一个模态的信息有助于另一个模态信息的处理，所以多模态的融合肯定是一个趋势。

再就是多机制，大家都说“感知还没有做好，怎么就开始去做认知了”，而事实上，感知和认知并不是独立的过程，不是说只有做好感知才能去做认知，实际上有些认知的东西能够反过来指导感知。我们做的很多工作，例如视觉问答或描述推理，这里面同时包含了认知和感知过程的，在其中加入推理机制往往能大幅提升模型的性能。不过现在很多工作还是还是单一机制，例如只加入注意或只加入推理。而人在感知或思考的过程中，事实上是多机制并行工作的。

黄铁军： 实际上我们最基本的感知，比如视觉，眼前能看到一幅清晰的图像，这个感觉本身就是认知过程加工的结果，而不是纯粹类似摄像头拍摄的结果。

王亮： 按照感知的定义来讲，视觉主要是采集形状信息、颜色信息等，偏向于表面多一点； 从图片中去理解包含什么事件，这更多属于认知的过程。

黄铁军： 所以视觉本身带有一些猜测和想像的过程，我们做梦闭着眼睛都能够想出栩栩如生的景象，就是认知在起作用。

王亮： 对。感知和认知不是一个阶梯的关系，其实是一个循环。

吴玺宏： 今天做视觉、听觉、语音、语言和机器人的几位老师讲的都不错，我也学习到很多。我们也想到多模态的整合，但怎么来整合？是浅显的整合还是深度的整合？

我的想法是首先不要着急做分类。我们现在的模式是，人费劲打标签，然后去做分类。但我们千万不要把感觉和知觉只看成分类，事实上重要的不是分类，而是如何表达——表达还没做好，就开始分类，是不对的，因为类别是人感知、认知之后来定的。

刚才山老师提到，怎么定义感知的问题。我这里借用维纳定义信息的话，他说“主体首先要有效地生存，之后才有信息”； 我借过来说，“主体首先要有效地生存才有智能”。智能的特点，首先应该是“适应性”，例如什么场合说什么话，如果只是靠人来定义类别，限定类别数，那智能体不可能延伸去识别重来没有出现的类别； 只有去提高它的适应性，它才能在适应的过程中，具备自动生成新概念、新类别的能力，从而解决大千世界中的识别问题。