南京大学
NANJING UNIVERSITY

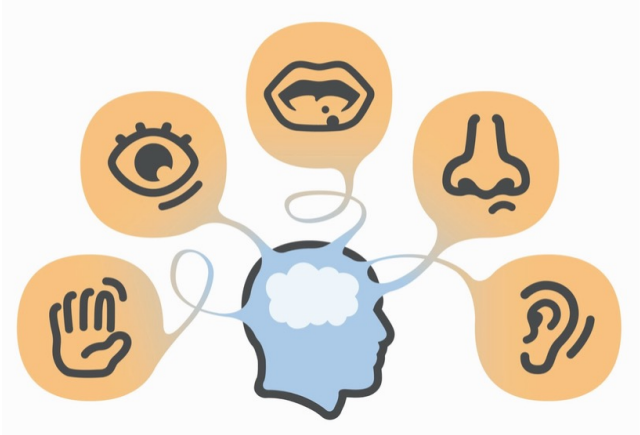
感知技术

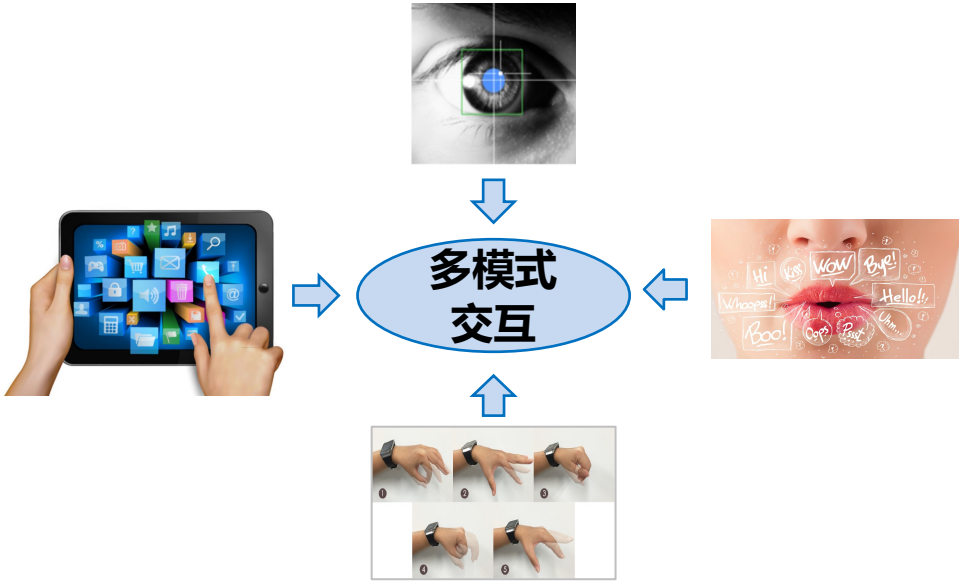
殷亚凤
智能软件与工程学院
苏州校区南雍楼东区225
yafeng@nju.edu.cn，https://yafengnju.github.io/

多模态感知技术

- **概述**
- 多模态输入
- 多模态融合

概述





单一模态的输入无法满足多变的真实场景需求，也不能有单一的交互模式适应于多任务。根据场景和任务智能选择合适的模态，甚至**组合多种模态**是高效、准确理解真实世界和用户意图的不二选择。

多模态感知技术

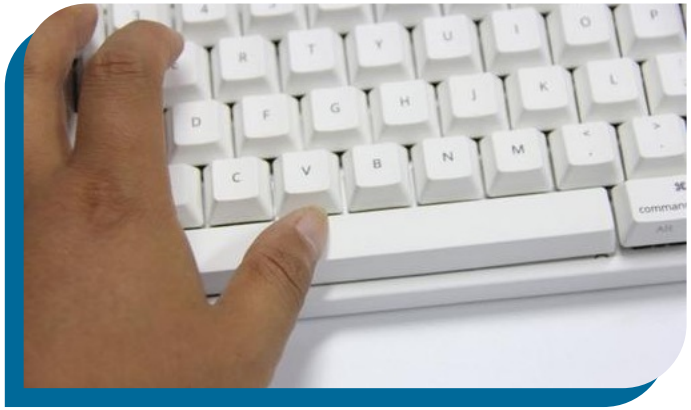
- 概述
- **多模态输入**
- 多模态融合

多模态输入

实际证明，单一模态的输入无法满足多变的真实场景需求。以下是**AR设备中常见的模态输入**：

AR多模态输入：键盘输入

- **传统键盘输入**



- 打字是一项与人体工程学息息相关的任务，是现代人生活中不可或缺的一部分。
- 我们通常花费大量时间在键盘打字上，对键盘输入方式的研究也日益增多。

AR多模态输入：键盘输入

- **虚拟键盘输入**

- 基于手持终端（手机和平台）的应用一般可以调用屏幕上的虚拟键盘。
- 这是对开发者最简单，也对用户最容易接受的方式。

AR多模态输入：键盘输入

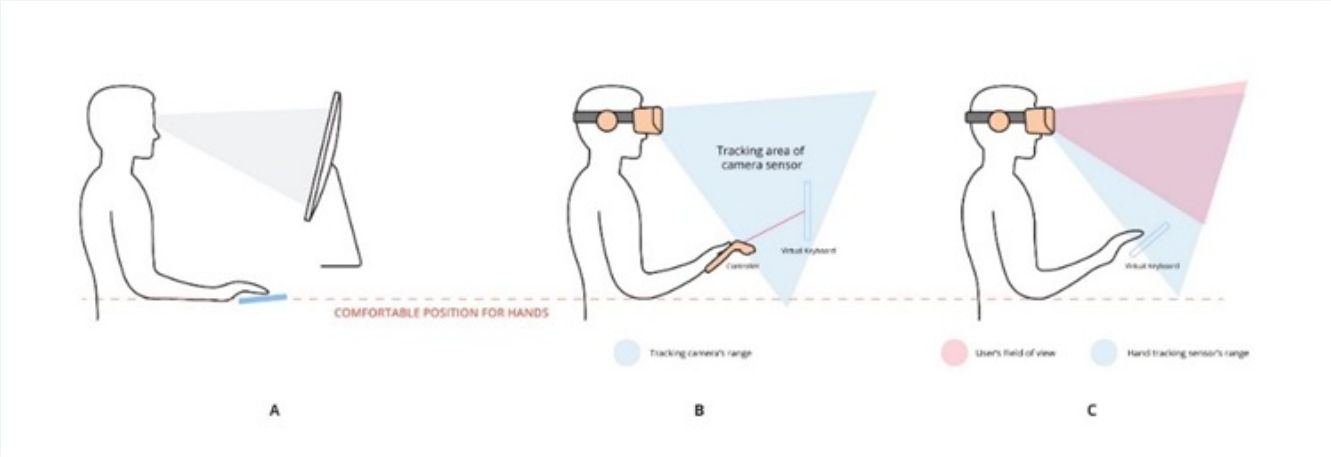
- **虚拟键盘输入**

增强现实输入方法相比原始的方法存在区别：

（A）手和显示器在标准的位置进行文本输入

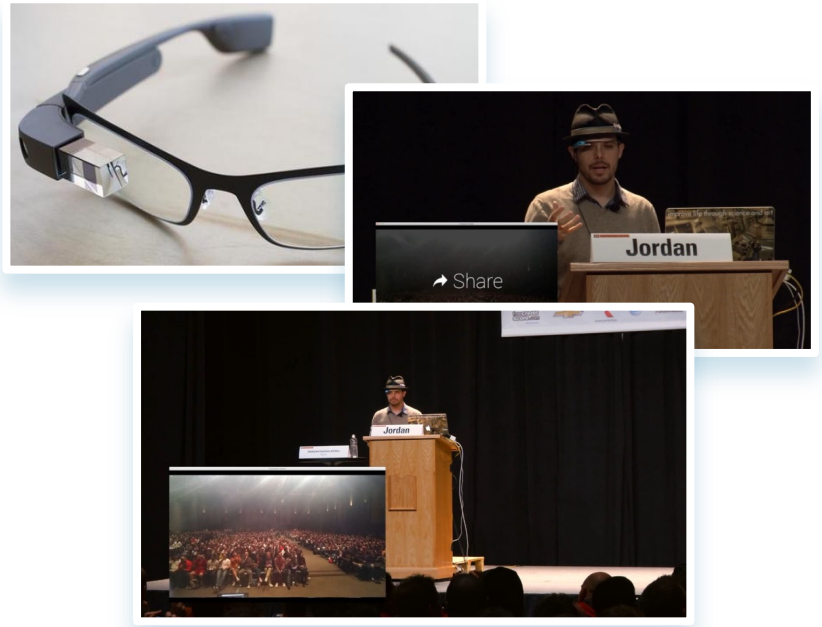
（B）使用VR设备时，用控制器进行文本输入

（C）使用VR设备时，使用手部光学追踪进行文书输入



AR多模态输入：键盘输入

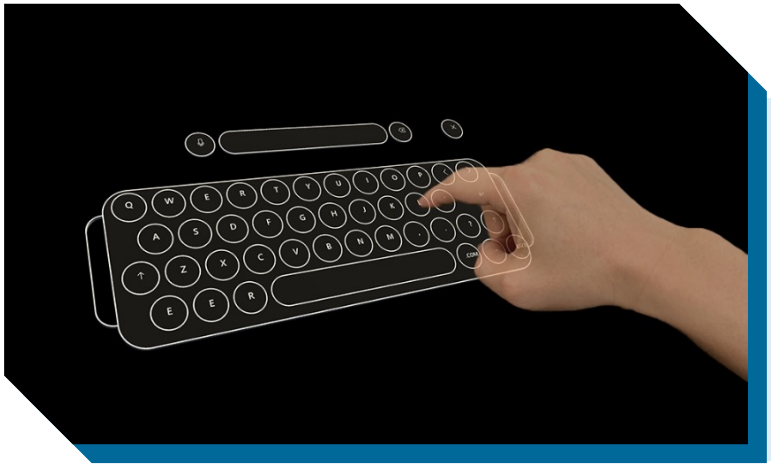
- **虚拟键盘输入**



Google Glass等专业设备则需要通过眼神、手势、控制器等方式将指针移动到目标键位上，再点击确认进行输入该字符，这种方式是非常缓慢的过程，需要花费大量的时间。

AR多模态输入：键盘输入

- **虚拟键盘输入**



手部跟踪键盘，是和实体键盘的输入方式是基本一致的，唯一的区别就是虚拟键盘代替了实体键盘，所以使用者可以不需要任何学习就可以直接使用。

AR多模态输入：键盘输入

- **虚拟键盘输入**

工作方式：通过HMD的前置摄像头捕获手掌和手势的位置。也就是说，使用者使用手掌在空中的位置来指示光标的位置，以作为基于手的“指向”。使用者根据他们的手在虚拟键盘上移动光标。

AR多模态输入：键盘输入

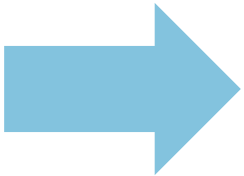
- **虚拟键盘输入**

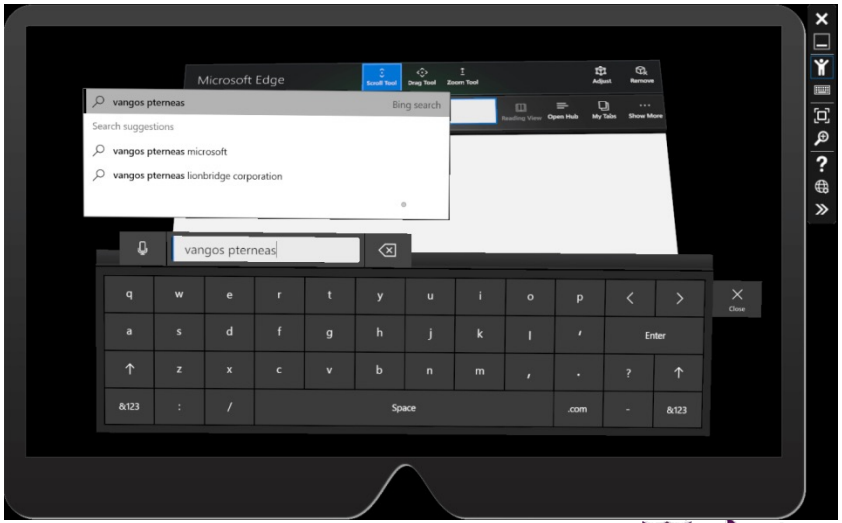
工作方式：光线在一个点处与键盘相交，并且出现指针作为提示。使用者要输入单词，需要使用光标将光标移动到相应的字母上。字母选择是通过点头动作通过点击头戴显示器的按钮进行的选择动作。

AR多模态输入：语音输入

- 在2017年，微软的技术可以让机器转录错误率以5.8%略优于人类水平5.9%。
- 微软已经在自己的增强现实设备上成功应用语音识别，如HoloLens。

语音识别
（听写识别）





AR多模态输入：语音输入

要启动语音输入，第一个步骤要实现**语音唤醒**，即通过特定词语启动增强现实设备，使其进入理解人类自然语言的模式。

AR多模态输入：语音输入

- **优点：**语音输入能够方便普通人，特别是部分有运动障碍的用户使用。

- 再者语音控制和输入可以有效减少使用者的操作时间，工作量，使用学习成本，符合人的原始习惯。这种方法能够有效让使用者有更好的体验。

AR多模态输入：语音输入

面临的挑战：

- 01 语音识别的准确性仍需要改进，口音，方言和小众语言仍然是语言识别所遇到的困难。
- 02 语音输入在公共场合不具有保密性，一些隐私无法通过语音输入。
- 01 对于背景音嘈杂的真实环境下，语音输入的准确率可能急剧下降，导致交互的失败。
- 04 目前的语言识别进行文本输入时，通常在输入之后，需要自己手工调整部分字句。
- 05 语音输入对于控制的细粒度无法准确表达。特别是对于缩放和移动等命令，语音输入无法准确量化程度。

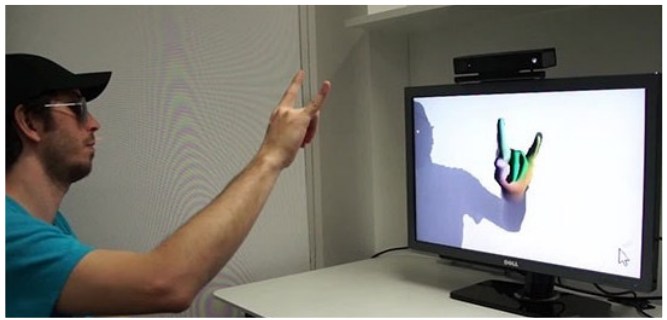
AR多模态输入：体感输入

- 体感输入是利用身体感觉作为输入模态的方式。
- 不局限于增强现实领域，体感输入的代表性技术是Kinect和Leapmotion。

AR多模态输入：体感输入

Kinect

Kinect是微软发行的Xbox的外接体感摄像机，通过**红外线发射器和红外线接收器来检测深度信息**。它具有即时动态捕捉、影像辨识的功能，可以识别使用者的肢体动作进行体感互动。



AR多模态输入：体感输入

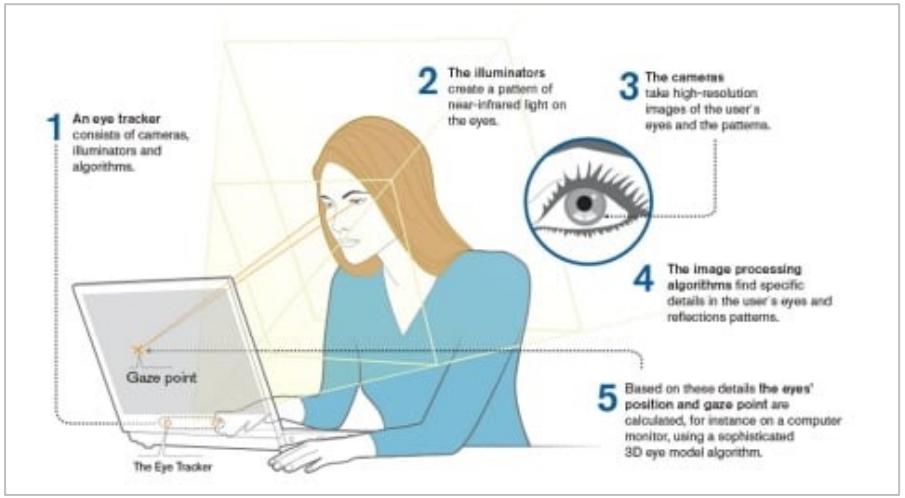
Leapmotion主要目标是跟踪手指运动，具有毫米级的精度，可以捕捉到绝大部分的手部动作。通过内置的两个不同角度的摄像头，捕捉两个方位的手部图像并重建出真实世界三维空间手部的运动信息。

AR多模态输入：眼神输入

眼动跟踪的基本原理是通过**摄像头捕捉眼睛反射的红外光来跟踪人的视线。**

使用者**通过移动目光来移动指针，再凝视或者眨眼来进行确定。**

重要目的：改善增强现实应用使用者体验。



AR多模态输入：眼神输入

1、眼动设备能够了解使用者在任何时间点将注意力都花在了哪里。

2、眼动设备的感知功能带来了大量有关使用者注意力的新数据源。 因此该信息资源是商业广告catch消费者眼球的重要法宝。

AR多模态输入：眼神输入

当前，眼动追踪的设备达到了一个全新的水平，如果只是PC端应用只需要一个眼动捕捉仪。

AR多模态输入

- **HoloLens**

  https://www.bilibili.com/video/BV1Rb411w7ff/

- **Google glass**

  https://www.youtube.com/watch?v=4EvNxWhskf8

多模态感知技术

- 概述
- 多模态输入
- **多模态融合**

多模态融合

- **传感器级融合**

优点	• 原始信息丰富,能提供其他融合层次所不能提供的详细信息, 精度最高。
缺点	• 所要处理的传感器数据量巨大,处理代价高,耗时长,实时性差;原始数据易受噪声污染,需融合系统具有较好的容错能力。
理论	• IHS变换, PCA变换,小波变换及加权平均等。
应用	• 主要应用多源图像复合、图像分析和理解。

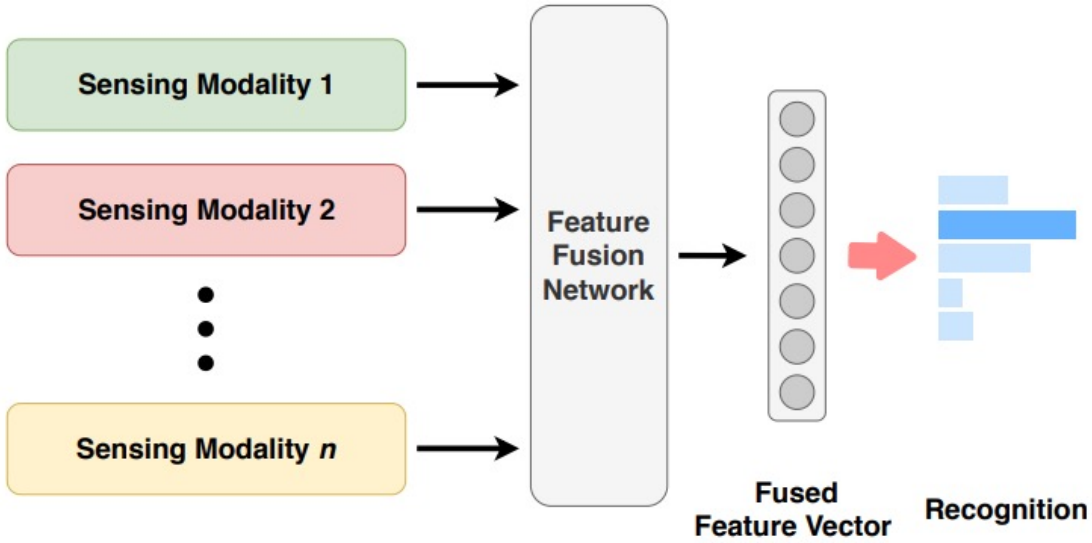
多模态融合

- **特征级融合**

优点	• 实现了对原始数据的压缩,减少了大量干扰数据,易实现实时处理,并具有较高的精确度。
缺点	• 在融合前要先对特征进行相关处理,把特征向量分类成有意义的组合。
理论	• 聚类分析法,贝叶斯估计法,信息熵法,加权平均法,D-S证据推理法,表决法及神经网络法等。
应用	• 主要用于多传感器目标跟踪领域,融合系统主要实现参数相关和状态向量估计。

多模态融合

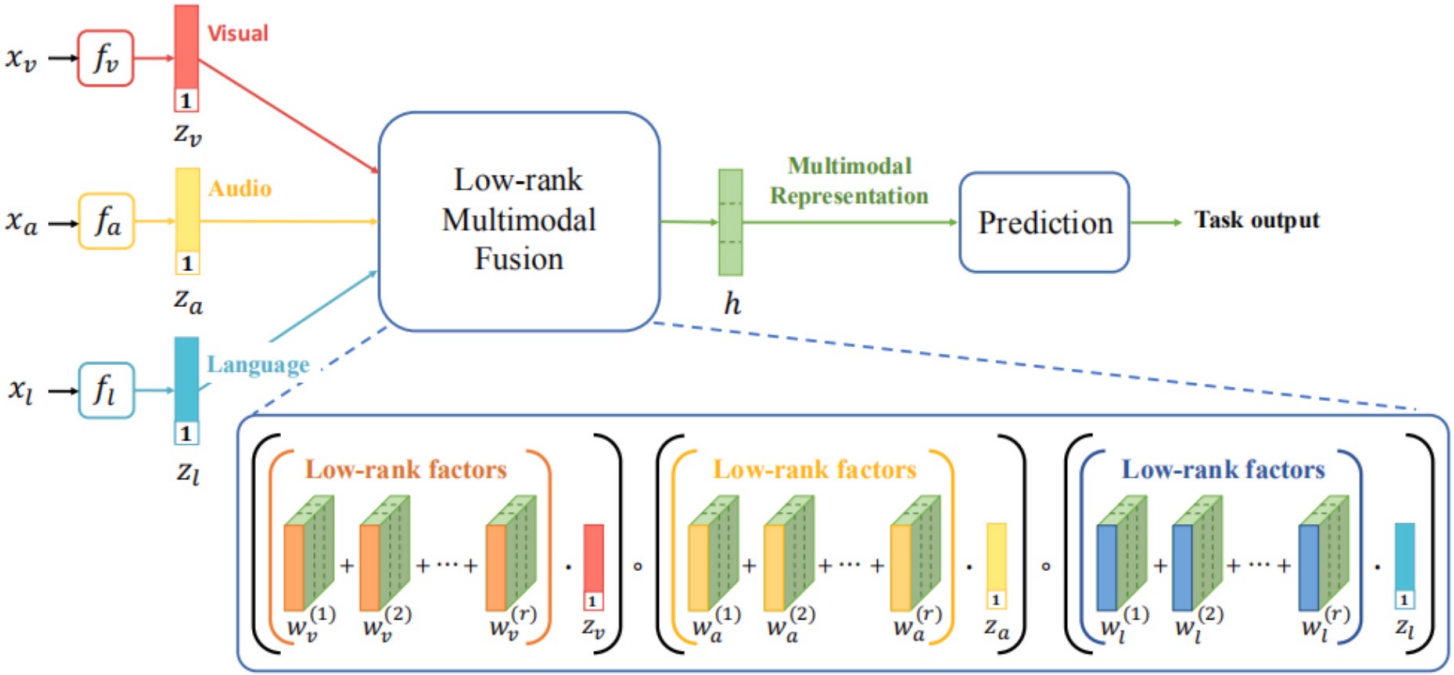
- 特征级融合



[1] Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. ACM Computing Surveys (CSUR). 2021 May 22;54(4):1-40.

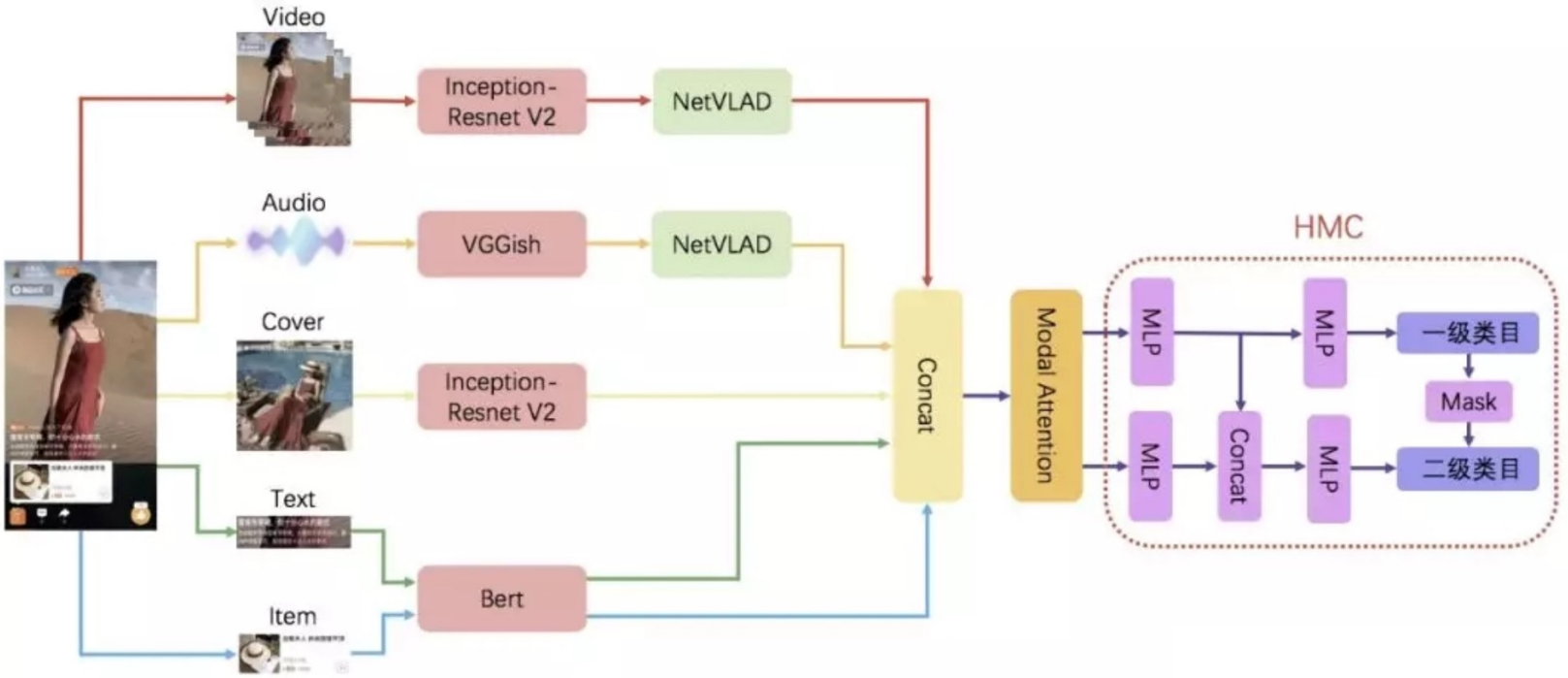
多模态融合

- **特征级融合**


x_v
f_v
Visual
1
z_v
x_a
f_a
Audio
1
z_a
x_l
f_l
Language
1
z_l
Low-rank Multimodal Fusion
h
Multimodal Representation
Prediction
Task output
Low-rank factors
$w_v^{(1)}$
$w_v^{(2)}$
$w_v^{(r)}$
1
z_v
Low-rank factors
$w_a^{(1)}$
$w_a^{(2)}$
$w_a^{(r)}$
1
z_a
Low-rank factors
$w_l^{(1)}$
$w_l^{(2)}$
$w_l^{(r)}$
1
z_l

多模态融合

- **特征级融合**


Video
Inception-Resnet V2
NetVLAD
Audio
VGGish
NetVLAD
Cover
Inception-Resnet V2
Text
Bert
Item
Concat
Modal Attention
HMC
MLP
MLP
一级类目
Mask
MLP
Concat
MLP
二级类目

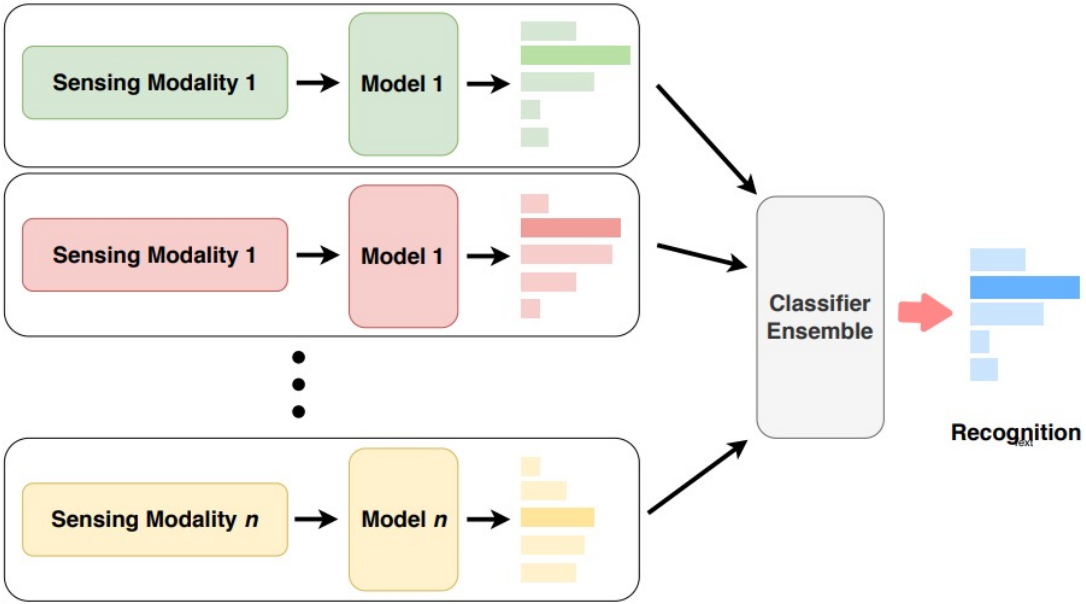
多模态融合

- **决策级融合**

优点	• 所需要的通信量小,传输带宽低,容错能力比较强,可以应用于异质传感器。
缺点	• 判决精度降低,误判决率升高,同时,数据处理的代价比较高。
理论	• 贝叶斯估计法、专家系统、神经网络法、模糊集理论、可靠性理论、逻辑模板法等。
应用	• 其结果可为指挥控制与决策提供依据。

多模态融合

- 决策级融合



[1] Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. ACM Computing Surveys (CSUR). 2021 May 22;54(4):1-40.

论文阅读报告——截止日期：10月28日晚23:59

- 提交方式：https://selearning.nju.edu.cn/（教学支持系统）

- 命名：学号+姓名+第*章。
- 若提交遇到问题请及时发邮件或在下一次上课时反馈。

论文阅读报告——截止日期：10月28日晚23:59

- **Human action recognition**

类一：Vision

[1] PeVL: Pose-Enhanced Vision-Language Model for Fine-Grained Human Action Recognition (CVPR 2024)
[2] Prompt-Guided Zero-Shot Anomaly Action Recognition using Pretrained Deep Skeleton Features (CVPR 2023)
[3] MoLo: Motion-augmented Long-short Contrastive Learning for Few-shot Action Recognition (CVPR 2023)
[4] Multi-Modality Co-Learning for Efficient Skeleton-based Action Recognition (MM 2024)

类二：Sensor

[1] AutoAugHAR: Automated Data Augmentation for Sensor-based Human Activity Recognition (UbiComp 2024)
[2] IMUGPT 2.0: Language-Based Cross Modality Transfer for Sensor-Based Human Activity Recognition (UbiComp 2024)
[3] Semantic Loss: A New Neuro-Symbolic Approach for Context-Aware Human Activity Recognition (UbiComp 2024)
[4] Spatial-Temporal Masked Autoencoder for Multi-DeviceWearable Human Activity Recognition (UbiComp 2024)

在上述某一个类别中任选2篇论文阅读，并撰写读书报告；报告格式采用软件学报模版（见课程主页），篇幅6-8页。

提问

Q & A

殷亚凤
智能软件与工程学院
苏州校区南雍楼东区225
yafeng@nju.edu.cn，https://yafengnju.github.io/

南京大学 NANJING UNIVERSITY